2017

Transcript Diversity In The Protozoan Parasite Toxoplasma Gondii

Maria Alejandra Diaz-Miranda
University of Pennsylvania, maleja.diaz@gmail.com

Transcript Diversity In The Protozoan Parasite Toxoplasma Gondii

Abstract
Technological advances have made possible to sequence RNA transcripts at unprecedented depth, enabling deep profiling of abundance and diversity under a variety of conditions. Such information permits refinement of draft genome annotation originally generated in the absence of transcript coverage data, and provides new insights into organismal biology and regulatory mechanisms. This dissertation provides an extensive analysis of mRNA-seq data from the obligate intracellular protozoan parasite Toxoplasma gondii, a ubiquitous pathogen of humans and other vertebrates. We produced and sequenced 24 strand-specific RNA libraries from several parasite strains and developmental stages, and examined these in con□junction with 45 additional mRNA-seq libraries produced by other groups.

The current reference genome annotation for T. gondii, generated using de novo methods informed by cDNA sequencing prior to mRNA-seq, identifies ~8300 protein-coding genes, fragmented by ~40K introns. Untranslated regions are incompletely defined, few alternatively-spliced transcripts are described, and non-coding transcripts remain largely unexplored. mRNA-seq datasets presented in this dissertation define a total of 2.7M introns, most observed at vanishingly low abundance. Using current annotation to define parameters minimizing false discovery yields ~60K likely splice junctions. Comparing the frequency of intron-spanning reads to the abundance of transcripts to which introns belong provides a reliable metric for estimating intron excision, readily distinguishing introns that are (i) universally used, (ii) alternatively-spliced, or (iii) likely insignificant. Genome-wide analysis suggests ~3000 annotated introns that should be deleted from the reference genome, ~1400 to be added as alternative isoforms, ~3100 as additions to existing annotation (often within UTRs) and ~3400 associated with novel transcripts. Transcriptomic expression is consis□tent with biological and phenotypic variation across the complex parasite life cycle, including undescribed differences in gene expression during intracellular tachyzoite replication. Strong circumstantial evidence also suggests that lncRNAs may play an important role in regulating stage-specific expression during sexual differentia□tion and sporogony. These results provide the basis for revising the reference T. gondii genome annotation available at ToxoDB.org and GenBank. Strategies developed in this dissertation also provide the basis for defining annotation criteria for other species, including related parasites responsible for malaria and conceivably other eukaryotes as well.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Biology

First Advisor
David S. Roos

Second Advisor
Scott Poethig

Keywords
Alternative splicing, Eukaryotic genome annotation, lncRNA, RNA-seq, Toxoplasma gondii, Transcriptomics

Subject Categories
Biology

TRANSCRIPT DIVERSITY IN THE PROTOZOAN PARASITE *TOXOPLASMA GONDII*

María Alejandra Díaz-Miranda

A DISSERTATION

in

Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

_____

David S. Roos PhD
E. Otis Kendall Professor of Biology

Graduate Group Chairperson

_____

Michael Lampson PhD
Associate Professor of Biology

Dissertation Committee

Brian D. Gregory PhD, Associate Professor of Biology

Zissimos Mourelatos MD, Professor of Pathology and Laboratory Medicine

Scott Poethig PhD, John H. and Margaret B. Fassitt Professor

Doris Wagner PhD, Professor of Biology

TRANSCRIPT DIVERSITY IN THE PROTOZOAN PARASITE *TOXOPLASMA GONDII*

María Alejandra Díaz-Miranda

ACKNOWLEDGMENT

I am immensely grateful to my advisor David S. Roos for welcoming me in his lab and guiding me throughout the scientific journey of a PhD program. David has been the best example of perseverance and patience, encouraged me not to give up and has always provided me with simultaneous liberty and support to explore different ideas. In his lab I was fortunate to meet people I admire personally and scientifically and that I consider I can call friends. I would particularly like to thank Daniel P. Beiting for showing me how everything was set up in the lab when I joined and for always being so keen to help. I am also extremely happy to have met Dinkorma Ouologuem, whom was always in the lab with a good laugh, good music and good advice about navigating the ups and downs of the graduate student life. The entire EuPathDB team at Penn assisted me with incalculable support to develop many of the ideas of my PhD project, and I am very grateful to have had the opportunity of interacting with everyone in the group. I would also like to thank all members of my thesis committee for being so understanding and for their concern about the development of my project, in particular Scott Poethig. Brian D. Gregory was especially supportive in my first years at Penn and I am grateful for all the help and input he has given me. Penn has also allowed me to live in Philadelphia, where I have met many incredible friends. Lorena, Ricardo, Paula, Lucero and Matias have been an incredible source of support and I am forever grateful for their friendship. Finally I would like to thank my family. My Dad Jaime, so organized, hard working and adventurous, has always encouraged me to explore the world. My Mom Camila, the best example of an independent woman with an incredible work ethic, was permanently there with comforting and caressing words, advice and laughs. My sister Maria Andrea,

who always makes me laugh and shares with me her passions of movies and new and interesting ideas about the world. And then Leo my husband and partner in life, who I met thanks to science, has seen me in my good and not so good days and has constantly supported me with his time and love. This accomplishment is yours too.

ABSTRACT

TRANSCRIPT DIVERSITY IN THE PROTOZOAN PARASITE *TOXOPLASMA GONDII*

María Alejandra Díaz-Miranda

David S Roos

Technological advances have made possible to sequence RNA transcripts at unprecedented depth, enabling deep profiling of abundance and diversity under a variety of conditions. Such information permits refinement of draft genome annotation originally generated in the absence of transcript coverage data, and provides new insights into organismal biology and regulatory mechanisms. This dissertation provides an extensive analysis of mRNA-seq data from the obligate intracellular protozoan parasite *Toxoplasma gondii,* a ubiquitous pathogen of humans and other vertebrates. We produced and sequenced 24 strand-specific RNA libraries from several parasite strains and developmental stages, and examined these in conjunction with 45 additional mRNA-seq libraries produced by other groups.

The current reference genome annotation for *T. gondii,* generated using *de novo* methods informed by cDNA sequencing prior to mRNA-seq, identifies ~8300 protein-coding genes, fragmented by ~40K introns. Untranslated regions are incompletely defined, few alternatively-spliced transcripts are described, and non-coding transcripts remain largely unexplored. mRNA-seq datasets presented in this dissertation define a total of 2.7M introns, most observed at vanishingly low abundance. Using current annotation to define parameters minimizing false discovery yields ~60K likely splice

junctions. Comparing the frequency of intron-spanning reads to the abundance of transcripts to which introns belong provides a reliable metric for estimating intron excision, readily distinguishing introns that are (*i*) universally used, (*ii*) alternatively-spliced, or (*iii*) likely insignificant. Genome-wide analysis suggests ~3000 annotated introns that should be deleted from the reference genome, ~1400 to be added as alternative isoforms, ~3100 as additions to existing annotation (often within UTRs) and ~3400 associated with novel transcripts. Transcriptomic expression is consistent with biological and phenotypic variation across the complex parasite life cycle, including undescribed differences in gene expression during intracellular tachyzoite replication. Strong circumstantial evidence also suggests that lncRNAs may play an important role in regulating stage-specific expression during sexual differentiation and sporogony. These results provide the basis for revising the reference *T. gondii* genome annotation available at ToxoDB.org and GenBank. Strategies developed in this dissertation also provide the basis for defining annotation criteria for other species, including related parasites responsible for malaria and conceivably other eukaryotes as well.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: INTRODUCTION & OVERVIEW

The promise of genomics to improve understanding of biology

It has been known for decades that much of the genetic potential of organisms, as encoded in their DNA, is put into practice through the action of proteins, which carry out a wide variety of structural and biochemical functions. This critical conversion is mediated by DNA *transcription* into messenger RNA (mRNA), and mRNA *translation* into protein … the 'Central Dogma' of molecular biology. More recently, we have come to appreciate that RNA molecules can also carry out a wide variety of regulatory functions, including the modulation of transcription, translation, and mRNA stability (Turner and Morris 2010; Raina and Ibba 2014; Radhakrishnan and Green 2016).

New technologies for nucleic acid library construction and sequencing allow us to profile the entire repertoire of organismal transcriptomes, under a variety of conditions, making it possible to improve our understanding of organismal biology by analyzing the abundance of mRNA and other RNA molecules. Technological developments have also made it possible to conduct similarly comprehensive analysis of other genomic datasets: chromatin modifications, protein composition, *etc.* The high degree of resolution made possible by these techniques reveals many previously unrecognized RNAs, posing a challenge: how to interpret this vast wealth of information? What is biologically meaningful, and what is meaningless noise?

The goal of this dissertation is to develop methods for extracting maximal information from modern transcriptomic datasets. How can we exploit mRNA transcriptomes to better define gene structures? What is the diversity of other RNA molecules? What role(s) do these RNAs play in regulating gene expression?

My dissertation research focuses on the protozoan parasite *Toxoplasma gondii*, an experimentally accessible human pathogen. Because *T. gondii* displays many of the molecular features observed in other eukaryotic species, it is likely that these methods will be broadly applicable to other organisms as well.

Eukaryotic gene expression

While the central dogma of molecular biology outlines the flow of information from DNA to RNA to protein, there are many twists to this basic story. For example, many RNAs – snRNAs, rRNA, tRNAs, and other small and long non-coding RNAs – are never translated into proteins. Instead, they perform biological functions by interacting with or recruiting protein complexes for specific functions. Because RNA is typically single stranded, it can fold onto itself by base-pairing complementary nucleotides within its sequence to create complex three-dimensional shapes capable of performing structural and catalytic functions (Li, Zheng, Ryvkin, et al. 2012; Li, Zheng, Vandivier, et al. 2012; Incarnato and Oliviero 2016). Modern genomic scale projects have revealed that up to three quarters of the human genome is transcribed in one or more cells … but only half of these transcripts appear to have protein coding potential (Djebali et al. 2012; Harrow et al. 2012). Which of these transcripts play functional roles remains an open question. Understanding transcript expression patterns, how gene expression is regulated, and the biological significance of previously uncharacterized transcripts, is likely to help us understand how cells and organisms alter their behavior in response to external signals.

One common mechanism of transcriptional gene regulation in eukaryotes involves *trans*-acting proteins that recognize and bind to *cis* regulatory sequences in DNA (promoters), typically upstream of the transcription start site. For example, eukaryotic

promoters often include a short sequence of T and A nucleotides ~25 nucleotides upstream of the transcription initiation site, that is bound by the TATA-binding protein. When transcription factors bind to a promoter, they recruit and/or modulate the activity of RNA polymerase on DNA (Todeschini, Georges, and Veitia 2014). After being brought to the promoter region by transcription factors, RNA polymerase starts synthesizing short fragments of RNA until it undergoes a conformational change, brought about by transcription factor-mediated phosphorylation of its C-terminal domain, thereby strengthening binding to DNA and forcing transcription factor dissociation (Wong, Jin, and Struhl 2014; Bowman and Kelly 2014). Cooperation between RNA polymerase and other proteins helps to initiate transcription. For example, activator proteins recognize DNA regulatory regions (enhancers), usually upstream of promoters, helping to recruit transcriptional machinery to the transcriptional start site via mediator proteins that communicate with the polymerase, favoring transcription (Plaschka et al. 2015).

Other biological structures may also affect transcription, including the histone complexes that help to organize the extremely long DNA molecules of eukaryotic chromosomes. Nucleosomal positioning with respect to promoters may impact transcription, as does how tightly DNA is wrapped around the nucleosome, which is determined by histone modifications (Gissot et al. 2007; Croken, Nardelli, and Kim 2013). Histone methylation, acetylation, and other modifications may block polymerase access (Todeschini, Georges, and Veitia 2014) or other components of the transcriptional machinery.

Transcript processing

Some non-coding RNAs are known to recruit chromatin-remodeling proteins to specific regions of the genome, modulating the addition of active or repressive marks onto histones (Rinn et al. 2007; Chu et al. 2015) and could also regulate the accessibility of the transcription machinery. Once transcription starts, elongation factors associate with the RNA polymerase and allow it to transcribe for long distances without dissociating from the DNA.

In eukaryotes, primary transcripts are typically processed co-transcriptionally, at the same time that RNA polymerase synthetizes the new transcript. This is possible thanks to the ability of RNA polymerase to change conformations after phosphorylation of its C-terminal domain, exposing other binding sites for multiple proteins that modify the structure of the RNA transcript as it is transcribed. Major modifications to RNA transcripts include the addition of a 5' 'cap', addition of polyadenine nucleotides in its 3' tail and removal of intervening sequences (introns) by the spliceosomal machinery. Terminal modifications act as protective measures providing transcript stability until translation is set to occur, and may be recognized by the translational machinery.

In contrast, intron excision affects the sequence of the mature mRNA, and often how it is translated into protein. The spliceosome, formed by small nuclear ribonucleoproteins (snRNPs) and auxiliary factors, recognizes specific dinucleotide sequences upstream and downstream of introns (donor and acceptor splice sites), as well as a branch site within intron boundaries, and it performs two trans-esterification reactions between the recognized sequences to excise introns from mRNA transcripts. Spliceosome recognition specificity drives alternative splicing of introns, choosing from

among the various combinations of strong and weak donor and acceptor splice sites within a transcriptional unit (Kornblihtt et al. 2013), providing yet another level of regulation that may yield mature mRNA variants (Wang et al. 2008).

Alternative splicing, based on differential use of splice donors and/or acceptors, may alter exon boundaries, include or exclude exons (exon skipping), or read through introns (intron retention). Because differential splicing affects internal mRNA nucleotide sequences, it may also affect sequence in any resulting translation products, particularly as nucleotide sequences are read in sets of three during translation into protein, so the addition or subtraction of nucleotides may shift the reading frame, altering all down-stream protein coding.

Mature mRNA transcripts may also be regulated at various levels, including RNA turnover mediated by degrading enzymes (Houseley and Tollervey 2009), and export from the nucleus where mRNAs are formed to the cytoplasm where translation occurs. Finally, it is also important to note that functional gene expression is also regulated at the level of protein translation and stability. Because of the relative ease of RNA trans-cript analysis using modern molecular biological techniques, however, steady-state transcript abundance is commonly used as surrogate for measuring functional gene expression.

Eukaryotic genomics, in the era of new sequencing technologies

Although the importance of DNA and RNA in protein production has been under-stood for more than 50 years, until the advent of high-throughput sequencing methods, this process could only be studied at the level of individual genes, mRNA, or proteins. Modern nucleic acid sequencing technologies make it possible to sequence billions of

nucleotides per day, enabling complete genome sequencing for essentially any organism, even for eukaryotes, whose genomes range from ~10^7-10^{11} bp.

The sequencing depth reached with modern sequencing technologies gives a near-complete snapshot of transcriptomes with a much higher dynamic range of expression levels and at a base-pair level resolution, in contrast to microarray technologies. However, the increase in depth in several NGS technologies, such as Illumina, SOLiD and 454, is possible thanks to the short length of reads that are sequenced (35-500bp). Short reads are assembled back into full-length transcripts by either *de novo* assembly, based on a reference genome (*ab initio* assembly) or with a combined approach merging these two strategies. The assembly process is not trivial, as it requires large memories in computing systems to assemble large quantities of short reads and also the ability to correctly assign sequencing reads to specific variants of the same transcript in the case of *de novo* assembly. Sequencing reads from both ends of cDNA fragments (100-250bp – paired end protocol) and joining the overlapping reads to form a longer read can help solve the problem posed by short reads. This strategy, however, will not solve assembly difficulties if the paired end reads don't overlap. An even better approach to avoid assembly problems is to sequence single RNA molecules or large fragments of near-complete cDNAs, as proposed by Helicos and Pacific Biosciences technologies. These technologies also have PCR amplification-free protocols, which usually results in better sequencing coverage of high GC regions, although their sequencing error rates are higher than those from short read technologies. An ideal strategy to study the transcriptome would be to combine sequencing results from different technologies as each of them have advantages and difficulties (Martin and Wang 2011)

The exploration of gene function and expression often begins with analysis of DNA sequence elements, along with attempts to define the most probable gene structure within a DNA sequence. *Ab initio* gene finders rely on recognizing structural features on genes that are extracted from the genomic sequence alone. Prokaryotic genome analysis is now relatively straightforward, for a variety of reasons. In addition to their relatively small genome size (typically 10^6-10^7 bp), prokaryotic promoter sequences are often highly conserved, enabling systematic identification. Further, the lack of intervening sequences fragmenting protein coding genes enables protein products to be predicted by simple translation of long open reading frames (ORFs) using the universal genetic code). In contrast, the promoters and regulatory sequences for most eukaryotic genes cannot currently be defined by specific sequence motifs. And splicing of primary mRNA transcripts produces protein coding regions or exons that are joined by non-coding regions or introns, greatly complicating prediction of complete coding sequences.

Sequencing technology may also be used to determine mRNA sequences after first converting RNA to DNA via the action of reverse transcriptase, and from the outset has been used to sequence individual clones mRNAs, or fragments of randomly-selected mRNAs (expressed sequence tags; ESTs). Because transcript abundance can differ by several orders of magnitude, however, comprehensive cataloging of transcripts initially relied upon hybridization to arrays constructed based on synthetic oligonucleotides, manufactured based on *ab initio* predictions of gene sequences. While immensely valuable for assessing gene-level transcript abundance, such methods are compromised by the inaccuracy of *ab initio* gene predictions, and usually unable to distinguish between alternatively-spliced transcripts.

With the continued increase in sequencing capacity (and decline in cost), it has recently become possible to consider sequencing at sufficient depth to effectively profile essentially all of RNAs in a cell or tissue. This strategy offers several advantages over array-based expression profiling. Because it does not depend on the accuracy of gene model predictions, it provides an assumption-free method for identifying transcripts, regardless of the polymerase used, transcript processing, or their potential to encode proteins. Further, deep sequencing permits identification of alternative transcription products from the same gene. Mapping mature mRNAs back to the reference genome provides direct evidence of alternative transcript initiation, termination, and/or splicing, providing the basis for improved gene-finding algorithms (*de novo* predictions), guided by mRNA sequences and other lines of evidence (DNA accessibility based on histone positioning, chromatin modification or nuclease sensitivity; transcription factor binding; etc). Methods have also been devised to address other concerns, such as the use of ribosomal profiling to identify transcripts that are actually loaded onto ribosomes, rather than steady-state mRNA levels that fail to take into consideration such important factors as transcript translatability or stability.

Integration of additional extrinsic features into gene finder algorithms permits better definition of gene models, including identification of alternatively spliced transcripts, definition of non-coding regions within genes or UTRs and identification of previously unannotated genes, including non-coding RNA transcripts. The challenge raised by this wealth of data, however, is how to distinguish between biologically important results, and rare events unlikely to manifest themselves in any significant way. For example, the first *ab initio* annotation of the human genome predicted approximately 30 thousand genes, but pilot projects designed to explore and annotate transcript diversity in detail suggests

up two 200 thousand distinct transcripts (Djebali et al. 2012; Harrow et al. 2012). Are these all functionally significant? The goal of this thesis is to exploit a model eukaryotic system, the protozoan parasite *Toxoplasma gondii*, to develop methods for improving the accuracy of genome annotation as well as applying this knowledge to enhance our understanding of organismal biology.

Biology of the protozoan parasite *Toxoplasma gondii*

Toxoplasma gondii is a unicellular eukaryotic parasite with interesting and clinically important biology. This species' developmental cycle encompasses both sexual and asexual stages, in distinct environments (Figure 1), presumably requiring distinct patterns of gene expression. Asexual *T. gondii* 'tachyzoites' are haploid obligate intracellular parasites, capable of invading any nucleated cell, in any warm-blooded vertebrate. Within the infected cell, tachyzoites establish a specialized 'parasitophorous vacuole' within which they replicate clonally, doubling every 8-16 hours until they lyse the infected cell 2-3 days later, releasing new tachyzoites that can infect neighboring cells. All pathogenesis by *T. gondii* (and the many thousands of other parasites in the phylum *Apicomplexa*) is directly attributable to rapid proliferation. Toxoplasmosis in humans is characterized by tissue damage, retinal degeneration, encephalitis, *etc.* In the absence of an effective immune response, the consequences can be fatal.

Tachyzoites are highly promiscuous, and readily able to cross epithelial barriers, including the blood/brain barrier and the placenta. During pregnancy, a primary infection in the mother can be transmitted to the developing fetus, leading to fetal abortion or severe congenital abnormalities. The underdeveloped immune system of the fetus is

unable to recognize and efficiently attack the rapidly dividing tachyzoites, leading to extensive tissue destruction.

In healthy adults, the immune system typically eliminates tachyzoites, but a subset of parasites differentiate into a slowly replicating 'bradyzoite' form that is poorly recognized by the immune system, and which can persist within specialized tissues cysts for the life of the host. Latent bradyzoite cysts periodically recrudesce to produce acutely lytic tachyzoites, providing a natural boost to the anti-*T. gondii* immune response … but also threatening immunocompromised individuals, including HIV/AIDS sufferers, or patients immunosuppressed for cancer chemotherapy or transplantation. Bradyzoite tissue cysts are particularly common in muscle tissues and the central nervous system, probably because they are able to persist in these locations for long periods of time as the immune response ramps up. Carnivory is probably the most common source of human infection, as bradyzoite cysts can be transmitted by ingestion of infected tissues.

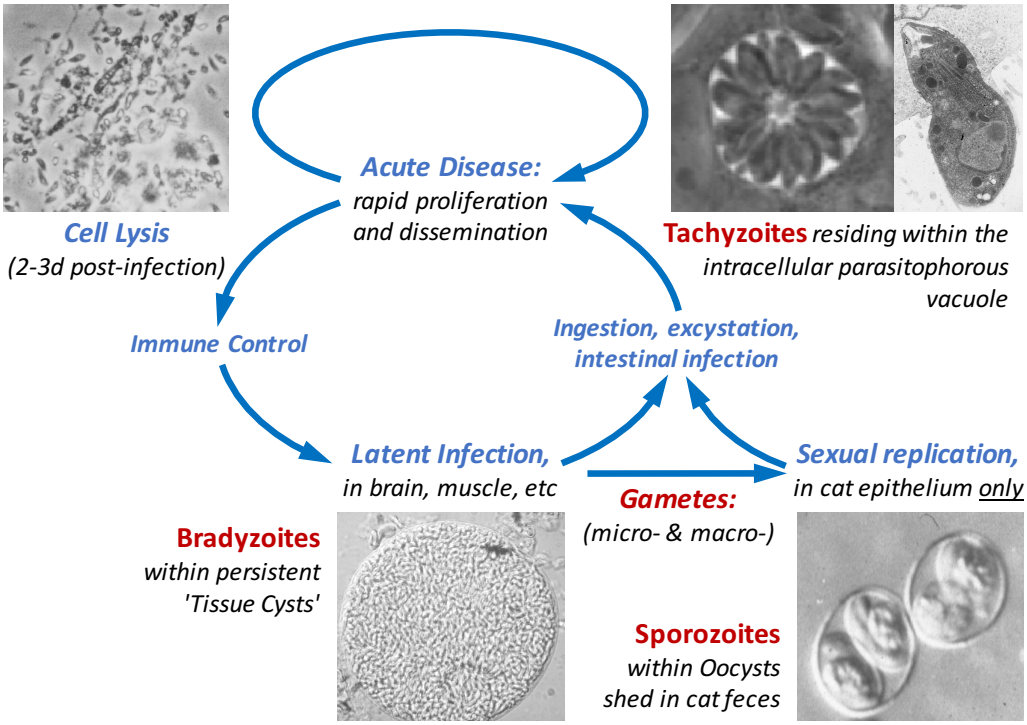


Figure 1. The complex development life cycle of the protozoan parasite *Toxoplasma gondii*.

Humans and other vertebrates become infected with *T. gondii* through the ingestion (or injection) of oocysts from the environment (contaminated soil or water) or tissue cysts from contaminated meat. Excystation is triggered by the acidic environment of the stomach, leading to the emergence of sporozoites or bradyzoites within the small intestine. These motile forms infect the intestinal epithelium and differentiate to form tachyzoites, which replicate rapidly, ultimately lysing the infected cell and invading neighboring cells and tissues throughout the body. This lytic cycle can continue indefinitely, and is responsible for all parasite pathogenesis. A small fraction of tachyzoites differentiate into tissue cysts, perhaps in response to immune attack, containing slow-growing bradyzoites, enclosed within a thin proteoglycan wall impervious to immune attack. Within the intestinal epithelium of felines (only), parasites undergo sexual differentiation to produce micro- and macro-gametes, which fuse to produce an oocyst that is released into the intestinal lumen and excreted in the feces. Upon exposure to oxygen and ambient temperature in the environment, these diploid oocysts undergo rapid meiotic reduction to produce haploid sporozoites.

The sexual stages of *T. gondii* only develop in feline intestinal epithelium, typically after ingestion of an infected mouse. Why sexual development doesn't occur in other

animals is unknown, but is typical of the phylum *Apicomplexa*. For example, *Neospora* parasites can also infect many species, but only replicate sexually in dogs. The *Plasmodium* parasites that cause malaria only undergo sexual differentiation in anopheline mosquitoes. In the feline intestinal epithelium, *Toxoplasma* tachyzoites differentiate into micro- and macro-gametocytes (analogous to sperm and eggs), and fuse to produce diploid oocysts that are shed in cat's feces. Upon exposure to ambient oxygen, meiotic reduction produces four haploid sporozoites, still encapsulated within the oocyst wall, but ready to start a new asexual cycle when the sporulated oocyst is ingested by the next host *via* contaminated soil or water. Sporulated oocysts remain viable in the environment for decades.

Once ingested by an uninfected (immunologically naive) host, bradyzoite tissue cysts or sporulated oocysts excyst in the gut, initiating a new round of invasion and asexual replication and completing the parasite life cycle. The ability to be transmitted *via* contaminated water or food (meat harboring bradyzoite tissue cysts, or vegetables grown in contaminated soil) probably accounts for the ubiquitous nature of this parasite. Approximately one third of the global human population is estimated to be infected with this parasite (Blader et al. 2015).

Although *Toxoplasma* is a diploid organism, capable of sexual reproduction, note that only haploid forms replicate. As a consequence, the successful progeny of sexual recombination (in cats) yields clonal parasites capable of rapid proliferation. The population biology of *T. gondii* is characterized by such clonal rockets, transmitted either via carnivory, or through selfing in cats (individual *T. gondii* zoites are totipotent, able to produce both micro- and macro-gametes). For example, while virtually all parasite

isolates show some evidence of relatively recent sexual recombination, most isolated from North America and Europe belong to just a few closely-related lineages (Howe and Sibley 1995; Khan et al. 2011; Lorenzi et al. 2016).

Experimental accessibility of *Toxoplasma gondii*

Consistent with their promiscuous nature, tachyzoite parasites can be cultivated indefinitely *in vitro*, in either transformed cell lines (HeLa, CHO, LM, MDBK, Vero, 3T3, *etc*) or primary cells (foreskin fibroblasts, astrocytes, macrophages, *etc*). Various stress treatments (oxidative stress, alkali shock, *etc*) can be used to induce bradyzoite differentiation *in vitro.* Moreover, infection of mice provides an *in vivo* model that closely mimics human disease. Sexual differentiation has not been reproducibly observed *in vitro*, but experimental infection of cats is a reliable (but expensive) method for carrying out traditional genetic crosses. The resulting progeny display typical Mendelian inheritance: four haploid sporozoites are produced from a diploid oocyst after meiotic division (Pfefferkorn, Pfefferkorn, and Colby, 1977; Dubey, Lindsay, and Speer 1998). Because only haploid forms replicate (see above), it is not possible to assess heterozygote phenotypes, but this limitation can be partially addressed by molecular genetic manipulation (see below).

An extensive experimental toolkit is available for exploring *T. gondii* molecular and cellular biology, including transfection vectors, strategies for insertional mutagenesis, selectable markers, conditional and inducible expression, knockout and gene manipulation systems (including the CRISPR technology), localization of organelles and/or genes by tagging with fluorescent probes, *etc* (Donald and Roos 1993; Sibley, Messina, and Niesman 1994; Roos et al. 1995; Soldati 1996; Belperron et al. 2001; Bradley, Li, and

Boothroyd 2004; Gubbels et al. 2008; Huynh and Carruthers 2009; Fox et al. 2009; Fox et al. 2011; Andenmatten et al. 2013; Rommereim et al. 2013; Sidik et al. 2016). Moreover, as a member of the protozoan phylum apicomplexan, including many other parasites of clinical and economic significance (*Plasmodium* species are responsible for malaria, *Cryptosporidium* is a leading cause of diarrheal disease in infants, *Eimeria* is the leading infectious disease concern in the poultry industry, *Babesia, Neospora* and *Theileria* are important cattle pathogens, *etc*), many of the experimental tools and insights gleaned from *Toxoplasma* have proved to be applicable to other parasites. For example, transfection systems developed for *Toxoplasma* provided the basis for transfection of *Plasmodium,* and insights into mechanisms of drug resistance, the process of cell invasion, the biology of the apicoplast (see below) and some aspects of host-parasite interactions have proved to be of general interest (Kim and Weiss 2004; Blader et al. 2015; McFadden and Yeh 2017).

Toxoplasma gondii cell and molecular biology

Toxoplasma harbors a typical complement of eukaryotic organelles, including a typical eukaryotic nucleus, cytoskeletal elements, complex endomembrane system, and mitochondria. Some organelles present distinctive features, however. For example, in contrast to the open mitosis observed in metazoan systems, the nuclear envelope remains intact throughout mitosis (closed mitosis, as observed in most unicellular eukaryotes). The nuclear envelope plays an important role in chromosome segregation, as it organizes the centrosome, mitotic spindle, and chromosomal centromeres to form the centrocone (Brooks et al. 2011; Farrell and Gubbels 2014). *Toxoplasma* also harbors various apicomplexan specific organelles, including the apicoplast (a non-

photosynthetic plastid acquired by secondary endosymbiosis of a eukaryotic alga, and retained as an essential organelle; (Köhler et al. 1997; Ralph et al. 2004), apical secretory organelles (micronemes and rhoptries), and the apical polar ring and conoid (Graindorge et al. 2016). These organelles constitute the apical complex, for which the phylum *Apicomplexa* is named. Secretory organelles play distinct roles in key aspects of the life cycle, including host cell attachment, invasion and egress (Besteiro et al. 2009; Huynh, Boulanger, and Carruthers 2014; Roiko, Svezhova, and Carruthers 2014; Huynh and Carruthers 2016), establishment and maintenance of the parasitophorous vacuole, and intracellular survival and interaction with the host cell (Saeij et al. 2007; Fentress et al. 2010; Reese et al. 2011; Fentress et al. 2012; Fleckenstein et al. 2012; Fox et al. 2016; Leroux et al. 2015).

At ~ 65 Mb (14 chromosomes), the *Toxoplasma* DNA genome is significantly smaller than animal, plant, and many other parasite genomes, but much larger than prokaryotic genomes. Genes are not arranged in operons, as in prokaryotic microbes, and transcription is mediated by typical eukaryotic polymerases and part of the basal transcriptional machinery, recruited by some conserved general transcription factors (TFIID, E, F, H, *etc*), under the influence of promoters, enhancers and chromatin marks similar to those found in other eukaryotes (Meissner and Soldati 2005; Sullivan et al. 2013). Apicomplexan transcription factors are dominated by AP2-integrase domain-containing proteins (Balaji et al. 2005), which are among the 20 most abundant Pfam domains in tissue-cyst forming coccidian parasites like *T. gondii* (Lorenzi et al. 2016).

Primary *T. gondii* mRNAs are capped and polyadenylated – although there is also evidence of non-polyadenylated transcripts – and contain numerous *cis*-introns, but no

known *trans*-splicing. Introns are excised using standard eukaryotic spliceosomal machinery, as human nuclear extracts can properly splice *T. gondii* primary transcripts, and vice versa. These features are important for both *ab initio* and *de novo* gene finders, as well-characterized human splice signals can be incorporated as intrinsic features during gene detection.

Toxoplasma genes were first cloned in the 1980s and sequencing of random expressed sequence tags (ESTs) in the 1990s defined several thousand additional genes expressed in tachyzoite parasites (Ajioka et al. 1998). Northern blotting, qPCR and analysis of EST frequency provided the first glimpse of transcript abundance, including identification of related transcripts suggesting alternative splicing (Donald et al. 1996; Chaudhary et al. 2005).

A first draft of the *T. gondii* reference genome completed in 2002 (shortly after completion of the human genome sequence) and assembled EST clusters mapped to the genome defined many gene models used to train *ab initio* and *de novo* gene finders predicting a total of approximately 8000 genes with protein coding potential. With the availability of annotated gene models, microarrays were designed to assess the transcriptional levels for all predicted genes (Bahl et al. 2010). This platform has been used by the global *T. gondii* research community to define stage- and strain-specific gene expression (*see* ToxoDB.org; Kissinger et al. 2003). As noted above, however, assessment of expression levels in microarray experiments poses several limitations. Background fluorescence signal corrections, typically excludes analysis of weakly expressed transcripts. Furthermore, since probe design is restricted to correct gene number estimation and requires preexisting knowledge of gene structure, microarray technology is

not well-suited to identifying errors in the present gene annotation, alternative transcripts, or unannotated genes. In contrast, RNA-seq technologies are able to profile all transcripts. These characteristics render RNA-seq technology suitable for evaluating transcriptome expression and improving the quality and accuracy of genome annotation.

Overview of this dissertation

By enabling comprehensive interrogation of organismal genomes, genomic-scale datasets are becoming an increasingly important tool in biological and biomedical research. This wealth of information holds promise for improved genome analysis, and new insights into the molecular mechanisms of many biological processes. The abundance of data also creates a serious challenge: How to better identify and interpret relevant information from these datasets? How do we discriminate biological significance and incorporate this information into decision-making? More specifically, is my gene-of-interest expressed? Is it correctly annotated, in all its forms?

The goal of this thesis is to develop analytic methods that permit discrimination of significance among genome-scale datasets, particularly transcriptome data derived from RNA-seq experiments. In pursuit of this goal, we have focused on the genome of *Toxoplasma gondii*, comparing current gene annotation to transcriptional evidence derived from RNA-seq experiments across all life cycle stages in order to identify errors in gene annotation, unannotated genes, and alternatively spliced transcripts.

As briefly described above, *T. gondii* is a unicellular eukaryotic parasite capable of causing disease in immunocompromised humans and animals, and a well studied organism, with a near complete genome sequence and a myriad of experimental tools. We constructed and sequenced 24 strand-specific RNA-seq libraries intended to explore

transcriptional variation across the intracellular lytic tachyzoite stage of the life cycle, and examined these together with 45 additional sequenced libraries profiling transcription in other life cycle stages. All of these datasets are now available on line at ToxoDB.org, and deposition in GEO is in progress.

These datasets have been used, in conjunction with other functional genomic information where appropriate, to explore the feasibility of identifying stage-specific markers based on transcriptional profiling data, and to improve the status of genome annotation, including untranslated regions, previously unrecognized exons and alternatively spliced transcripts, and novel transcripts, including long non-coding RNAs (lncRNAs). Throughout this dissertation research, I have attempted to emphasize methods that should be broadly applicable to other eukaryotic microbes for which similar datasets are available.

CHAPTER 2: USING SEQUENCING TECHNOLOGIES (AND OTHER LARGE-SCALE DATASETS) TO ASSESS AND IMPROVE GENOME ANNOTATION IN *TOXOPLASMA GONDII* (Adapted From Paper)

New sequencing technologies have made many species accessible to genomic-scale analysis, raising the challenge of how to integrate such information from various sources, discriminate biological significance, and make these results accessible to diverse end-user communities. We have exploited strand-specific RNA-seq analysis to profile the transcriptome of human host cells infected with the protozoan parasite *Toxoplasma gondii,* a prominent eukaryotic microbial pathogen responsible for disease during congenital infection and in immunosuppressed individuals (Tenter, Heckeroth, and Weiss 2000).

At ~65 Mb in length, the *T. gondii* genome is relatively compact (Lorenzi et al. 2016), but harbors most of the complexity described for other eukaryotic genomes, including ~8300 protein coding genes, ranging in length from <1 to >60 kb (ave ~4.8 kb), and fragmented by introns (range 0-60+; ave ~5.8) that follow consensus eukaryotic sequence constraints (primary *T. gondii* transcripts are properly spliced by human nuclear extracts). Extensive population genetic and functional genomic datasets are available for *T. gondii,* including additional RNA-seq data and other transcriptional profiles for various strains and developmental stages, proteomics data, chromatin marks, *etc* (ToxoDB.org; Gajria et al. 2008). As noted above (Chapter 1), numerous tools are also available for experimental manipulation of *T. gondii* in the laboratory (Roos et al. 1995; Sibley et al. 2002; Kim and Weiss 2004; Meissner et al. 2007; Sidik et al. 2016). We have generated strand-specific RNA-seq datasets for various *T. gondii* tachyzoite

strains, at multiple time points during their 48 hr intracellular *in vitro* replicative cycle, and analyzed these in parallel with datasets from other life cycle stages, to assess the accuracy of current genome annotation, identify new genes (including alternatively spliced transcripts), and assess stage-specific transcript expression and regulation (Chapter 3).

Methods

Parasite cultures, RNA isolation, RNA library construction and sequencing

T. gondii tachyzoites from four different strains (ME49, VEG, RH, GT1), were maintained by serial passage in human foreskin fibroblast (HFF) monolayers as previously described (Roos et al. 1995), infecting confluent monolayers with ~10^7 tachyzoites. For time course experiments, parasites were propagated in Vero cells (*Cercopithecus aethiops*) for two passages immediately before the infection of HFFs for RNA isolation, to avoid inadvertent contamination with human material from the previous infectious cycle. After media removal, cell monolayers were scraped in 700ul of Qiazol, and RNA isolated from cell lysates using the Qiagen miRNEasy mini kit, according to the manufacturer's instructions. Six biological replicates were collected per strain and timepoint, and checked for RNA quality using a BioAnalyzer (Agilent). Strain M4 bradyzoites, strain CZ-H3 enterocytes (gametocytes), and strain M4 oocysts were prepared and RNA isolated as previously described (Buchholz et al. 2011; Fritz, Buchholz, et al. 2012; Juránková et al. 2013; Basso et al. 2013; Hehl et al. 2015). Biological replicates were pooled, and total and polyA+ selected RNA used to construct strand-specific mRNA (and in some cases small non-coding RNA) libraries as previously described (Li, Zheng, Vandivier, et al. 2012; Elliott et al. 2013), and sequenced on Illumina Hi-Seq 2000 (see

Table 1 for total number of reads per library). All RNA-seq data described in this study are available from the *Toxoplasma* Genome Database, at ToxoDB.org (Gajria et al. 2008), along with other RNA-seq datasets and diverse additional information (Fritz, Bowyer, et al. 2012; Minot et al. 2012; Reid et al. 2012; Lorenzi et al. 2016).

Alignment of RNA-seq reads to the *T. gondii* genome: the ToxoDB pipeline for mapping RNA-seq reads

For transcript assembly, RNA-seq reads were initially mapped onto *T. gondii* ME49 genome release 28 using RUM (Grant et al. 2011); subsequent studies used GSNAP (Wu and Nacu 2010) to map to genome version 29. The RUM alignment pipeline takes advantage of the speed of Bowtie (Langmead et al. 2009) to map against both the genome and transcriptome; unmapped reads are then mapped against the genome using Blat (Kent 2002), and Information from all three mappings is then merged. GSNAP was configured to look for both known and novel splicing. Coverage was determined for unique and non-unique alignments (separated by strand when possible). Strand orientation of splicing was determined based on the usage of GT/AG, GC/AG, or AT/AC dinucleotide pairs on the plus strand (or their complements on the minus strand). In cases where strand could not be defined, the program applies a probabilistic splice model to determine orientation (Wu and Nacu 2010). Performance for both RUM and GSNAP tools is comparable to the best RNA-seq alignment tools available at the time this study was completed (Engström et al. 2013).

Algorithm for gene model learning and prediction

Gene model training and predictions were performed using a version of CRAIG (A. Bernal et al. 2007) that integrates RNA-seq data, encoded as features derived from the

mapping of reads to the *T. gondii* genome, as described above. The evidence integration strategy and feature encoding for RNA-seq data have been reported previously (Bernal, Crammer, and Pereira 2012; Bernal and Pereira 2012). ToxoDB v28 gene annotations were used for training, after filtering to exclude genes with evidence of significant alternative splicing (see below). The learned model integrates *ab initio* features such as segment length distributions, with features derived from junction-spanning and coverage reads. We sought completeness in transcript prediction by forcing CRAIG to define at least one transcript model for each non-overlapping transcript junction (putative intron) with read counts >3, and for overlapping junctions, those displaying >20% of the highest support observed in any overlapping junction within the same gene model.

Assessment of genome annotation, analysis of alternative splicing and visualization

To assess the quality of the reference *T. gondii* genome annotation, constructed largely based on *ab initio* methods informed by EST sequences from tachyzoite stage parasites only, expression data was retrieved for all predicted introns in 69 RNA-seq samples available in ToxoDB.org (Table 1), including samples from many strains, and most *T. gondii* life cycle stages (see Chapter 1 for a description of the parasite's asexual and sexual life cycles). This yielded a list of 2,731,523 candidate introns, but many were observed in only one or two samples, or at very low abundance levels in any sample. Introns observed <6 times overall, or with <3 read in any of the 69 samples considered, were excluded from further analysis, leaving a total of 147,715 for in-depth analysis (including 997 previously-annotated introns not satisfying the above criteria).

Table 1. List of *T. gondii* RNAseq datasets used in this study (ToxoDB release 28)

	Ref	Strain	Stage	Host	Cond	Time	RNA	Str Spec	Ins Size	Read Ln	Total Reads *	*Tg* Unique *	% †	Total ISRs ‡	%
Intracellular Tachyzoites -- Diverse Strains															
GT1 Sibley	*1*	1	GT1	Tachyzoit	HFF cells	*In vitro*	72 hr?	polyA+	No	unknown	100+100?	164,728,174	28,962,965	17.6%	2,630,466	9.1%
ME49 Sibley	*2*	1	ME49	"	"	"	"	"	"	unknown	100?	31,824,251	22,950,642	72.1%	2,726,503	11.9%
ARI	*3*	2	ARI	Tachyzoit	HFF cells	*In vitro*	72 hr?	polyA+	No	220	40+40	127,295,324	20,890,634	16.4%	1,607,718	7.7%
B41	*4*	2	B41	"	"	"	"	"	"	"	"	39,158,742	7,627,165	19.5%	465,197	6.1%
B73	*5*	2	B73	"	"	"	"	"	"	"	"	27,010,286	6,775,092	25.1%	446,632	6.6%
BOF	*6*	2	BOF	"	"	"	"	"	"	"	"	31,918,787	8,692,004	27.2%	593,508	6.8%
CAST	*7*	2	CAST	"	"	"	"	"	"	"	"	30,684,829	12,688,027	41.3%	772,439	6.1%
CASTELLS	*8*	2	CASTELLS	"	"	"	"	"	"	"	"	25,943,434	11,269,884	43.4%	698,953	6.2%
CEPdelta	*9*	2	CEPdelta	"	"	"	"	"	"	"	"	22,911,358	9,730,742	42.5%	646,186	6.6%
COUGAR	*10*	2	COUGAR	"	"	"	"	"	"	"	"	26,336,254	10,043,346	38.1%	630,228	6.3%
DEG	*11*	2	DEG	"	"	"	"	"	"	"	"	24,488,338	13,221,779	54.0%	791,392	6.0%
FOU	*12*	2	FOU	"	"	"	"	"	"	"	"	27,511,865	4,801,559	17.5%	345,916	7.2%
GPHT	*13*	2	GPHT	"	"	"	"	"	"	"	"	17,408,230	4,098,107	23.5%	298,438	7.3%
GT1	*14*	2	GT1	"	"	"	"	"	"	"	"	30,487,790	11,530,934	44.2%	919,285	6.8%
GUYDOS	*15*	2	GUYDOS	"	"	"	"	"	"	"	"	27,575,023	13,468,205	85.4%	765,443	3.3%
GUYKOE	*16*	2	GUYKOE	"	"	"	"	"	"	"	"	45,820,315	23,550,991	51.4%	1,347,240	5.7%
GUYMAT	*17*	2	GUYMAT	"	"	"	"	"	"	"	"	26,315,451	9,347,030	35.5%	600,355	6.4%
MAS	*18*	2	MAS	"	"	"	"	"	"	"	"	19,686,294	10,123,378	51.4%	611,495	6.0%
ME49	*19*	2	ME49	"	"	"	"	"	"	"	"	25,200,682	11,045,627	43.8%	653,377	5.9%
P89	*20*	2	P89	"	"	"	"	"	"	"	"	27,368,978	16,818,388	61.5%	1,034,813	6.2%
PRUdelta	*21*	2	PRUdelta	"	"	"	"	"	"	"	"	35,047,626	19,656,333	56.1%	1,140,965	5.8%
RAY	*22*	2	RAY	"	"	"	"	"	"	"	"	23,593,872	9,785,081	41.5%	616,145	6.3%
Rhdelta	*23*	2	Rhdelta	"	"	"	"	"	"	"	"	61,603,537	29,353,242	47.6%	1,824,046	6.2%
ROD	*24*	2	ROD	"	"	"	"	"	"	"	"	45,907,606	16,795,259	36.6%	1,315,830	7.8%
RUB	*25*	2	RUB	"	"	"	"	"	"	"	"	44,678,085	17,164,186	38.4%	1,114,844	6.5%
TgCATBr44	*26*	2	TgCATBr44	"	"	"	"	"	"	"	"	51,063,437	17,750,373	34.8%	1,118,740	6.3%
TgCATBr5	*27*	2	TgCATBr5	"	"	"	"	"	"	"	"	22,403,348	6,857,924	30.6%	413,230	6.0%
TgCATBr9	*28*	2	TgCATBr9	"	"	"	"	"	"	"	"	63,264,704	17,443,532	27.6%	1,372,374	7.9%
VAND	*29*	2	VAND	"	"	"	"	"	"	"	"	41,694,538	28,878,303	69.3%	1,846,138	6.4%
VEG	*30*	2	VEG	"	"	"	"	"	"	"	"	18,166,749	9,144,480	50.3%	598,049	6.5%
WTD3	*31*	2	WTD3	"	"	"	"	"	"	"	"	34,270,722	14,811,235	43.2%	937,868	6.3%
Intracellular Tachyzoites -- Developmental Series																
Reid D3	*32*	3	VEG	Tachyzoit	HFF cells	*In vitro*	72 hr	polyA+	No	200-250	76+76	65,238,810	51,911,446	79.6%	4,096,024	7.9%
Reid D4	*33*	3	"	"	"	"	96 hr	"	"	"	"	77,988,674	63,623,584	81.6%	5,170,913	8.1%
RH 2 hr	*34*	4	RH	Tachyzoit	HFF cells	*In vitro*	2 hr	polyA+	Yes	275-375	100	1,894,365	194,671	10.3%	35,875	18.4%
RH 22 hr	*35*	4	"	"	"	"	22 hr	"	"	"	"	14,927,266	8,930,207	59.8%	1,605,429	18.0%
RH 36 hr	*36*	4	"	"	"	"	36 hr	"	"	"	"	16,839,750	15,348,364	91.1%	2,209,144	14.4%
GT1 2hr	*37*	4	GT1	Tachyzoit	HFF cells	*In vitro*	2 hr	"	"	275-375	"	3,918,897	364,309	9.3%	64,074	17.6%
GT1 4hr	*38*	4	"	"	"	"	4 hr	"	"	"	"	833,498	111,155	13.3%	23,529	21.2%
GT1 8hr	*39*	4	"	"	"	"	8 hr	"	"	"	"	2,453,961	514,210	21.0%	99,994	19.4%
GT1 16hr	*40*	4	"	"	"	"	16 hr	"	"	275-375	"	59,500,308	21,584,941	36.3%	3,712,516	17.2%
ME49 2hr	*41*	4	ME49	Tachyzoit	HFF cells	*In vitro*	2 hr	"	"	55-150	"	60,290	4,250	7.0%	480	11.3%
ME49 4hr	*42*	4	ME49	Tachyzoit	HFF cells	*In vitro*	4 hr	"	"	"	"	59,182,474	5,850,571	9.9%	523,097	8.9%
ME49 8hr	*43*	4	"	"	"	"	8 hr	"	"	"	"	42,147,632	5,833,363	13.8%	479,805	8.2%
ME49 16hr	*44*	4	"	"	"	"	16 hr	"	"	"	"	19,467,179	4,387,649	22.5%	403,446	9.2%
ME49 36hr	*45*	4	"	"	"	"	36 hr	"	"	275-375	"	19,313,474	10,951,950	56.7%	1,763,689	16.1%
ME49 44hr	*46*	4	"	"	"	"	44 hr	"	"	"	"	13,648,696	12,332,981	90.4%	1,722,462	14.0%
VEG 2 hr	*47*	4	VEG	Tachyzoit	HFF cells	*In vitro*	2 hr	"	"	55-150	"	116,612,403	13,005,698	11.2%	1,088,000	8.4%
VEG 4 hr	*48*	4	"	"	"	"	4 hr	"	"	"	"	58,188,707	7,680,123	13.2%	661,484	8.6%
VEG 8 hr	*49*	4	"	"	"	"	8 hr	"	"	"	"	57,496,732	9,139,881	15.9%	742,082	8.1%
VEG 16 hr	*50*	4	"	"	"	"	16 hr	"	"	"	"	26,968,837	6,543,629	24.3%	665,172	10.2%
VEG 36 hr	*51*	4	"	"	"	"	36 hr	"	"	275-375	"	22,772,838	12,418,256	54.5%	2,058,765	16.6%
VEG 44 hr	*52*	4	"	"	"	"	44 hr	"	"	"	"	14,768,871	8,826,301	59.8%	1,369,143	15.5%
Bradyzoite development																
Knoll acute mouse	*53*	5	ME49	Tachyzoit	HFF cells	*In vitro*	10 days	polyA+	No	unknown	50+50?	619,510,492	868,532	0.1%	51,040	5.9%
Knoll chronic mouse	*54*	5	"	Bradyzoit	Mouse	*In vivo*	28 days	"	"	"	"	662,065,868	1,441,169	0.2%	100,103	6.9%
InVitro Bz	*55*	9	ME49	Bradyzoit	HFF cells	*In vitro*	7 days	polyA+	Yes	55-150	50	29,120,884	26,411,628	90.7%	2,098,534	7.9%
InVivo Bz	*56*	8	M4	Bradyzoit	Mouse	*In vivo*	21 days	"	"	275-375	100	113,123,601	27,297,107	24.1%	4,452,108	16.3%
Gametocyte development																
Hehl Tz	*57*	9	CZ-H3	Tachyzoit	HFF cells	*In vitro*	control	polyA+	Yes	unknown	100+100?	199,141,200	85,515,887	42.9%	10,548,750	12.3%
Hehl D3	*58*	9	"	Gametocy	Cat intesti	*In vivo*	3 days	"	"	"	"	552,571,698	12,947,597	2.3%	1,719,245	13.3%
Hehl D5	*59*	9	"	Gametocy	"	"	5 days	"	"	"	"	195,225,948	81,478,584	41.7%	11,566,841	14.2%
Hehl D7	*60*	9	"	Gametocy	"	"	7 days	"	"	"	"	751,346,454	128,435,940	17.1%	18,013,378	14.0%
Oocyst development																
Oocyst D0	*61*	6	M4	Oocyst	NA	unsporulat	control	polyA+	Yes	55-150	58	22,416,214	10,037,743	44.8%	815,928	8.1%
Oocyst D4	*62*	6	"	"	"	sporulated	4 days	"	"	55-150	50	20,628,790	18,970,427	92.0%	1,508,317	8.0%
Oocyst D10	*63*	6	"	"	"	"	10 days	"	"	55-150	50	20,243,335	18,524,507	91.5%	1,427,680	7.7%
Other samples																
SR3 uninduced	*64*	7	RH cSR3	Tachyzoit	HFF cells	*In vitro*	control	polyA+	No	unknown	100+100?	102,451,076	94,223,358	85.7%	11,359,466	12.1%
SR3 4hr induced	*65*	7	"	"	"	SR3-induct	4 hr	"	"	"	"	89,992,092	82,711,487	91.9%	9,793,519	11.8%
SR3 8hr induced	*66*	7	"	"	"	"	8 hr	"	"	"	"	91,194,210	83,917,767	92.0%	10,126,194	12.1%
SR3 24hr induced	*67*	7	"	"	"	"	24 hr	"	"	"	"	96,979,952	87,825,880	90.6%	10,539,187	12.0%
ncRNA RH	*68*	9	RH	Tachyzoit	HFF cells	*In vitro*	33 hr	ncRNA	Yes	15-45	38	37,861,116	2,061,028	5.4%	127,575	6.2%
ncRNA ME49	*69*	9	ME49	"	"	"	24 hr	"	"	"	"	34,330,519	2,064,798	6.0%	145,050	7.0%
											Total (all samples):	**5,573,795,740**	**1,469,567,425**		**153,771,851**	
											excluding 'Other':	**5,120,986,775**	**1,116,763,107**		**111,680,860**	
											Unigue intron junctions:				**2,731,523**	
											High quality strand-specific libraries (green):	**2,354,385,967**	**506,589,822**		**63,627,926**	

References

1. Lorenzi H *et al., Nature Communications;* 2016 (ref 7)
2. Minot S *et al., Proc Natl Acad Sci;* 2012 (ref 18)
3. Reid AJ *et al., PLoS Pathogens;* 2012 (19)
4. This study
5. Pittman KJ *et al., BMC Genomics;* 2014 (20)
6. Fritz HM *et al., PLoS1;* 2012 (17)
7. Yeoh LM *et al., Nucl Acids Res;* 2015 (21)
8. Buchholz KR *et al. Eukaryot Cell;* 2011 (51)
9. Unpublished; available from ToxoDB.org

** read fragments for paired-end libraries (to avoid couble-counting); denominator for FPKM calculations*

† low % unique mapping reads is attributable to host cell RNA; cf ME49 time course in rows 42-46)

‡ total # intron-spanning reads (ISRs); denominator for ISRPM calculations

Table 1. List of T. gondii RNAseq datasets used in this study (ToxoDB release 28). Green highlighting indicates 20 high quality samples used define the prevalence of alternative splicing and mechanisms of transcriptional regulation; pink highlighting indicates reasons for exclusion of other samples from the analyses presented in Chapter 2; see text for further discussion.

Because strand-mapping information was not retained in the ToxoDB pipeline implementation of RUM, strandedness was assigned by analyzing five nucleotides up- and downstream of each intron to determine the most probably splice donor and acceptor. Analysis of the abundance distribution of dinucleotides pairs for each intron, showed that (as expected) the most common splice signal (on the plus strand) was 5'-GT/AG-3', which is >80 times more abundant than any of the other possible 63 intron combinations. 5'-GC/AG-3' and 5'-GA/AG-3', were the next most common (enriched 1.8 & 1.4 times, respectively). We therefore further filtered this intron list to include only introns that contained the major splice signal 5'GT/AG3'. This procedure yields a total of 66,104 introns for examination in greater detail.

Once strand was recovered, introns were assigned to gene structures to determine those fully contained within a previously-annotated gene model, those lying fully within intergenic regions in the draft annotation (potential extensions of existing gene models, or associated with previously unrecognized genes), and those overlapping draft gene models. For introns associated with a specific gene, reads spanning that intron should be comparable to reads mapping to the mature transcript. The abundance of intron-spanning reads (ISRs) per million reads in the library (ISRPM) was therefore plotted as a function of the number of reads mapping to the assigned gene, normalized to gene size (FPKM = read fragments per kilobase of transcript, per million total mapped reads).

Preliminary analysis using all 69 RNA-seq samples in Table I revealed poor correlation between intron and gene expression for a some introns, invariably attributable to poor quality samples in which relatively few reads could be mapped to the parasite (or host) genome. These samples were therefore excluded from analysis of splice junction usage, as were samples for which only non-strand-specific reads are available, and samples involving splicing machinery mutants (pink shading in Table1). Note, however, that data from all these samples remains available in ToxoDB, and were reviewed after the completion of our analysis of high confidence introns.

All further analysis was conducted using a final set of 59,755 introns, from 20 samples representing all parasite life cycle stages, in multiple strains (green shading in Table 1). To analyze the prevalence of alternative splicing, the abundance of each intron was compared to its most abundant alternative(s), if any. Data was visualized using DataGraph 4.1 (Visual Data Tools; Figs 2.4-2.6 & 2.12-2.13).

Results

Transcriptional insights from RNA-seq, applied to *Toxoplasma gondii*

Prior to the development of high throughput methods for mRNA sequencing, eukaryotic gene finding relied upon on *ab initio* methods (predicting gene structure based on primary sequence alone), and *de novo* strategies informed by cDNA sequences from expressed sequence tags (ESTs; Burge and Karlin 1998; Salamov and Solovyev 2000; Wei and Brent 2006). The much greater coverage provided by low cost RNA-seq methods greatly enhances gene model accuracy, however, enabling the identification of previously unrecognized transcripts, refinement of untranslated region (UTR) annotation,

definition of stage- and strain-specific transcripts, recognition of alternative splice junctions, *etc* (Trapnell et al. 2010). When available, additional genomic-scale datasets (chromatin marks, transcription factor binding sites, protein expression data, *etc*) can also be exploited further improve the accuracy of gene model prediction (Lamesch et al. 2012).

*Tg*HXGPRT was the first alternatively-spliced gene identified in the protozoan parasite *Toxoplasma gondii* (Donald et al. 1996), based on the presence or absence of an 'exon skip' polymorphism encoding a 49 amino acid insertion including an acylation motif responsible for protein association with parasite membranes (Chaudhary et al. 2005). The reference model in the official GenBank annotation (and ToxoDB) includes this exon-skip polymorphism as exon III (Fig 2.1 track 1; *blue* indicates transcription from left-to-right, *i.e.* on the forward, or top strand). Numerous ESTs map to *Tg*HXGPRT (track 2), confirming both splice variants: exon III is missing from eleven ESTs (HXGPRT-I), but included in six (HXGPRT-II). Western blotting indicates similar relative abundance of HXGPRT-I *vs* HXGPRT-II at the protein level (Chaudhary et al. 2005). This well-validated example of an alternatively-spliced gene was used as a positive control to define and assess parameters for alternative transcript identification genome-wide.

RNA-seq data provide vastly greater experimental support: average depth for the experiment presented in Fig 2.1 is >700 reads (track 3; ~$2^{9.5}$ on the log scale plot shown in track 5). The most common apparent transcript initiation site (in this steady-state analysis) occurs at ~6,795,950 (*heavy magenta arrow*), ~100 nt downstream of the annotated 5' end. In keeping with common conventions from the pre-RNA-seq era, this gene was originally annotated based on the longest, rather than the most abundant

cDNA clone. Log-scale representation reveals the range of 5' ends identified by RNA-seq, although we cannot exclude the possibility of 5' exonuclease activity or premature termination during reverse transcription. Most transcripts appear to terminate close to the annotated 3' end (*filled magenta circle*), although alternative low abundance termination sites are also evident, most prominently ~500 nt downstream (*open magenta circle*).

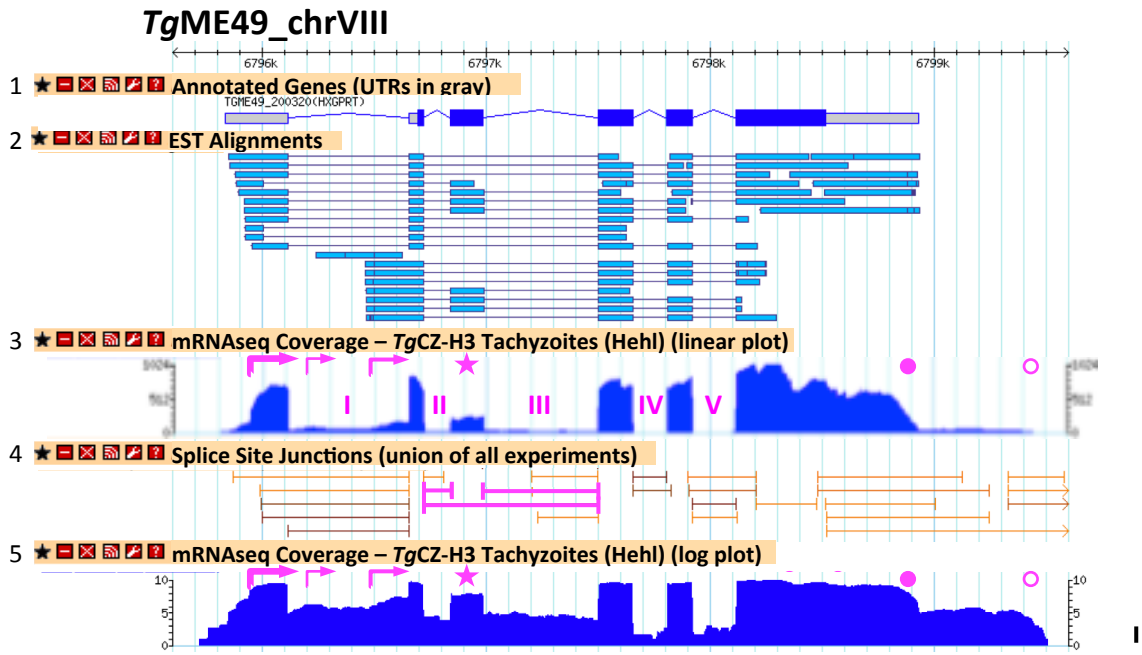


Figure 2.1. Reading RNA-seq data.

The annotated *T. gondii* HXGPRT gene (TgME49_200320), aligned to EST sequences and RNA-seq data (both linear & log coverage plots), including intron-spanning reads (brackets). Note the presence of multiple 5' ends (at steady-state), likely corresponding to multiple promoters (magenta arrows), of which only the longest and most prominent would permit excision of intron I (which lies within the 5'UTR). EST coverage in intron I was previously misinterpreted as intron read-through variants (Donald et al. 1996). Magenta brackets highlight the well-validated exon skip variant (star) responsible for membrane association of HXGPRT isoform II (Donald et al. 1996; Chaudhary et al. 2005). Multiple (low abundance) 3' ends are also observed (circles).

Six well-defined exons are clearly identifiable in the RNA-seq coverage plots, corresponding precisely to the annotatation *Tg*HXGPRT-II. Exon III (*magenta star*) is less abundant than the other five, however (seen most clearly in the linear representation; track 3), providing evidence of alternative splicing, consistent with ESTs, Northern blotting, protein immunoprecipitation, proteomics and protein structure data (ToxoDB.org and (Chaudhary et al. 2005)). RNA-seq reads that map across intron junctions (intron-spanning reads; ISRs) are indicated by horizontal brackets in track 4 (pooled information from numerous experiments). *Magenta brackets* highlight introns corresponding to the known HXGPRT exon-skip polymorphism: for the experiment shown (*Tg*CZ-H3 tachyzoites), 149 reads could be unequivocally mapped to intron II, 187 reads span intron III, and 180 reads span introns II+III (excising exon III), defining the HXGPRT-I transcript. Pooling all available experimental data (from multiple samples) provides overwhelming support for this exon-skip polymorphism (8573, 11475, and 12785 reads, respectively; ToxoDB release 28).

Intron-spanning reads (ISRs) also identify numerous unannotated introns, but these are significantly less abundant. For example, while intron I is supported by 391 reads in the experiment shown (27K in all studies), an alternative splice donor 112 nt upstream is supported by 7 ISRs and 227 ISRs respectively, and this alternative intron is also supported by EST data; two ISRs (61) support yet another alternative donor 8 nt further upstream. Note that none of these alternatives affects the predicted protein sequence, however, as intron I lies within the 5' untranslated region (UTR). Alternative ISRs mapping to the HXGPRT coding sequence also seem unlikely to be biologically meaningful, as the most common (87 reads in all studies, but none in the experiment shown) extends

intron IV by 17 nt, which would introduce a translational frame shift and premature termination eliminating most of the phosphoribosyl transferase domain (Pfam00156).

In addition to defining the exon skip polymorphism distinguishing HXGPRT-I & II, the original cDNA clones were interpreted to suggest retention of intron I in some transcripts (Chaudhary et al. 2005). Read coverage within intron I is significantly higher than other introns (~10% exon depth *vs* <5% for other introns), but careful examination of the RNA-seq data reveals that coverage is non-uniform, displaying gradually increasing depth in the direction of transcription, suggesting alternative promoters (*lighter magenta arrows*). This interpretation is consistent with both EST evidence and review of the original cDNA clones. Transcript initiation within intron I would of course preclude intron excision.

In sum, using the highly-curated *Tg*HXGPRT gene as a positive control for reanalysis based on RNA-seq data improves the definition of UTRs, confirms exon boundary annotation and a known exon-skip variant, permits identification of additional rare (and probably biologically meaningless) splice variants, and reveals that the previously-described intron retention is more likely attributable to alternative transcript initiation within intron I. Applying such analyses genome-wide offers the prospect of significantly improved gene model definition. For example, as shown in Fig 2.2, previous annotation (ToxoDB release 7.3, produced without the benefit of RNA-seq data) failed to define 5' UTRs for 5777 of the 8323 protein-coding genes in the reference *T. gondii* annotation, and 3' UTRs for 4859 (*gray bars*). Incorporating RNA-seq information now identifies 7219 5' UTRs and 7296 3' UTRs, with a modal 5' UTR length of ~750 nt, and 3' UTR length of ~500, similar to HXGPRT, above (*blue* in Fig 2.2).

Fig 2.3 presents an expanded genomic region, extending upstream of HXGPRT (the right-most gene in this panel), including RNA-seq evidence from both tachyzoite and gametocyte (enteroepithelial) stage parasites (Hehl et al. 2015), along with additional genomic-scale data from chromatin immunoprecipitation studies (Gissot et al. 2007). These experiments support the reference annotation of *Tg*ME49_200310 (immediately to the left of HXGPRT), which is heavily transcribed in tachyzoites (tracks 5,6,8), gametocytes (track 10), and other life cycle stages (not shown). Low abundance unannotated ISRs never exceed 2% of annotated intron abundance for this gene. Chromatin activation marks (H3K4me3 & H3K9ac; track 1) are consistent with a 242 nt region mediating divergent transcription of *Tg*ME49_200310 & 200320 (HXGPRT).

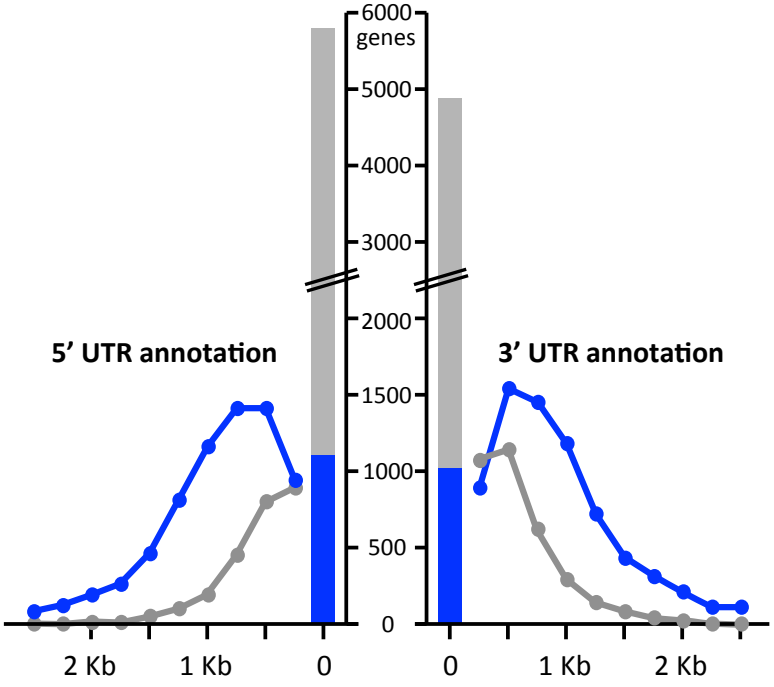


Figure 2.2. Length distribution of annotated UTRs. Lengths of UTRs before (*gray*) and after (*blue*) incorporating RNA-seq data into gene finding algorithms.

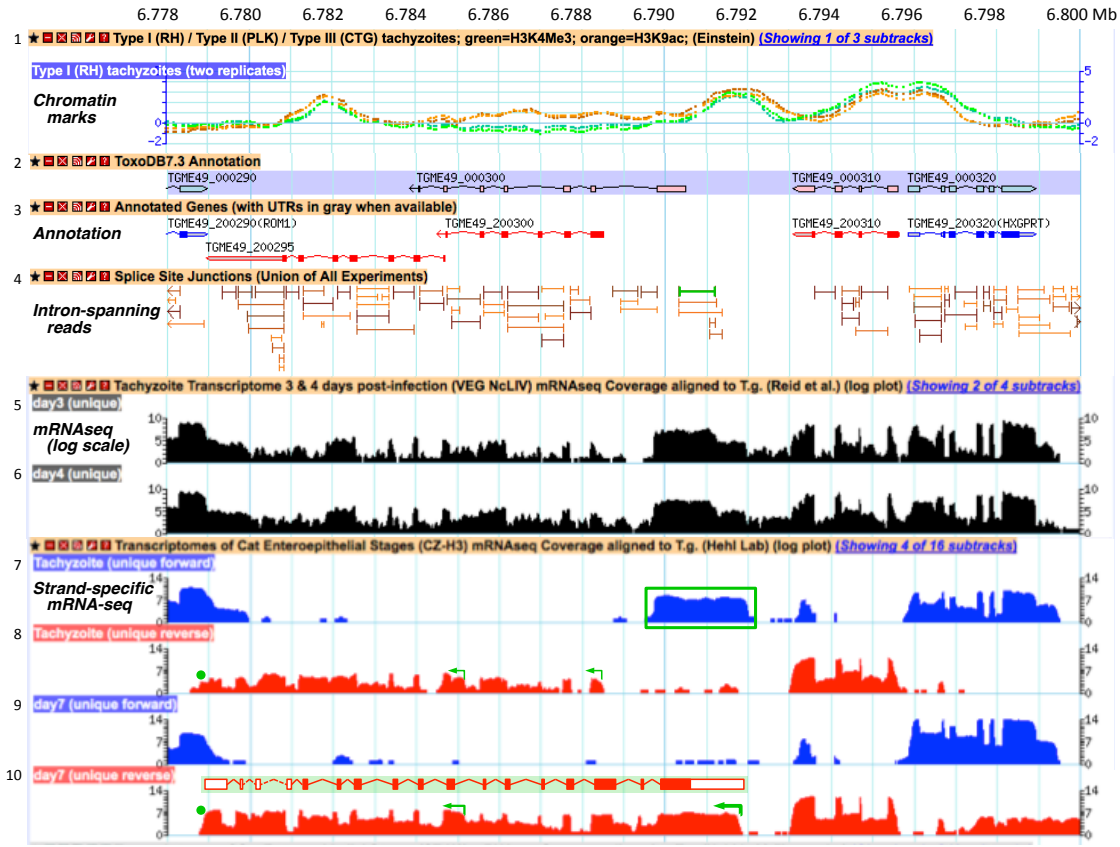


Figure 2.3. Strand-specific mRNA-seq reveals a multitude of alternative splice variants and antisense RNAs, including long non-coding RNAs.

Genome browser view of a 22kb region (*Tg*ME49_chrVIII: 6778-6800kb) displaying publicly available chromatin marks (H3K4me3 & H3K9ac, track 1), annotated gene models (including HXGPRT, at right, track 3), non strand-specific mRNA-seq data (tracks 5,6), and selected strand-specific mRNA-seq data from this study (tracks 7-10). Color intensity in track 4 indicates 10-fold differences in the abundance of junctional reads (dark brown > light brown > orange). *Arrows* and *circles* (tracks 8,10) indicate likely 5' and 3' termini, respectively; *green shadowed* transcript (track 10) proposes a corrected gene model; the *green box* highlights a likely long noncoding RNA (lncRNA). See text for further discussion.

In contrast, RNA-seq data argues for revised annotation of the adjacent region (ChrVIII: 6.778-6.792 Mb). *Ab initio* gene-finding methods previously suggested a seven-exon transcript (*Tg*ME49_000300; track 2), while more recent annotation taking

(limited, non-strand-specific) tachyzoite-stage RNA-seq data into account were used to define two transcripts: *Tg*ME49_200300 & 200295 (track 3). Strand-specific sequencing of multiple life cycle stages helps to reconcile these discordant views. In tachyzoite stage parasites (track 8), steady-state RNA levels indicate two 5' ends, consistent with the above gene models (*light green arrows*), but provide no support for either exon I or intron I from the first draft annotation (*Tg*ME49_000300). In gametocyte stages, however (track 10), transcription initiates far upstream (*heavy green arrow*), at a position consistent with chromatin modification data (track 1), and similar to the original *ab initio* predictions. The resulting 17-exon gene model (*green shading*) encompasses all exons previously suggested by any annotation, along with additional exons revealed by high-depth sequencing (mostly within the 3'UTR). As noted for the HXGPRT gene, while CDS intron excision is generally efficient, relatively high transcript abundance is sometimes observed within UTR introns, due to a combination of alternative transcript initiation, termination, splicing, and/or inefficient intron excision.

Interestingly, while the first exon of this gene (from ~6.790-6.792 Mb) is only observed in gametocyte stage parasites (track 10), this region is heavily transcribed on the opposite (forward, *blue*) strand in tachyzoites (track 7). The lack of an open reading frame on this strand, and lack of chromatin marks typically associated with active Pol II-mediated mRNA transcription (track 1) suggests that it is a long non-coding RNA (lncRNA). This interpretation is consistent with the low prevalence of introns mapping to this transcript:

Just 39 reads were observed spanning the most abundant intron seen in this experiment (97 in all experiments), at least 4-fold fewer than expected based on transcript

abundance (compare with read coverage troughs corresponding to introns for the HXGPRT gene, for example). This lncRNA is missing in gametocytes, *i.e.* it's presence is inversely correlated with transcription on the reverse (*red*) strand.

Genome-wide analysis of putative introns, and alternatively splicing

In order to exploit the utility of RNA-seq data for improving genome annotation, including the identification of stage-specificity, UTRs, alternative splicing, *etc,* we generated several strand-specific libraries from various parasite strains, at different times after infection of human host cells *in vitro* (*see* Methods), and considered the resulting RNA-seq data in parallel with all other *T. gondii* RNA-seq datasets available in the ToxoDB database (release 28). The 69 samples summarized in Table 1 (above) derive from nine studies involving various parasite strains, mutants, life cycle stages and host species.

A total of 1.1 x 10^9 mRNA-seq reads from wild-type parasites map uniquely to the *T. gondii* (not host) genome, and define ~2.7M splice junctions. The vast majority of these correspond to the U2 (GT/GC-AG) or U12 (GA-AG) splice consensus, suggesting genuine spliceosomal origin, but >90% were observed at very low abundance (<6 total reads, or <3 in the most abundant sample), and were not analyzed further. Introns whose strand could not be unambiguously determined were also excluded from this study. Further restricting analysis to strand-specific libraries with >10^6 parasite-specific reads leaves a total of 20 samples – 11 of which were generated specifically for this study, (and 16 of which are presented here for the first time). These samples provide support for >97% of the ~40K introns annotated in the current reference genome, in addition to ~20K unannotated introns that might reasonably be considered for inclusion in the reference annotation. Adding back ~1K annotated introns unsupported by RNA-seq

evidence yields a total of 59,755 candidate introns for further analysis (see Supplemental File 1).

Assuming uniform read coverage across all full-length transcripts, unambiguous mapping, and efficient splicing … and ignoring alternative splicing for the moment, one would expect a linear relationship between the abundance of intron-spanning reads (ISRs) and all reads that map to a transcript, normalized for transcript length (Venables et al. 2008; Katz et al. 2010). All introns were mapped to individual genes or intergenic regions, and intron abundance was plotted as a function of transcript abundance, as shown in Figure 2.4. Intron abundance (ISRPM) is defined for each junction as the number of ISRs spanning that junction, normalized to the total number of ISRs mapped in that experiment (in millions). Transcript abundance (FPKM) is defined for each gene as the number of reads (or read pairs) mapping to that gene, normalized to gene length (in kilobases) and the total number of read (pairs) mapped in that experiment (in millions).

The data in Fig 2.4 are derived from a single experiment on the acutely lytic (tachyzoite) stage of *T. gondii* strain CZ-H3 (Basso et al. 2013; Juránková et al. 2013). Annotated introns are shown in black, while unannotated introns are shown in gray; unannotated intergenic introns are not displayed, as their FPKM values are not readily defined. Two populations of introns are immediately evident: low abundance introns mapping to genes at various expression levels (shaded area in Fig 2.4), and introns that appear to be efficiently excised (*i.e.* ISRPM proportional to FPKM, regardless of expression level; diagonal in Fig 2.4). It is clear that the latter are predominantly annotated (*black*), while the former are predominantly unannotated (*gray*).

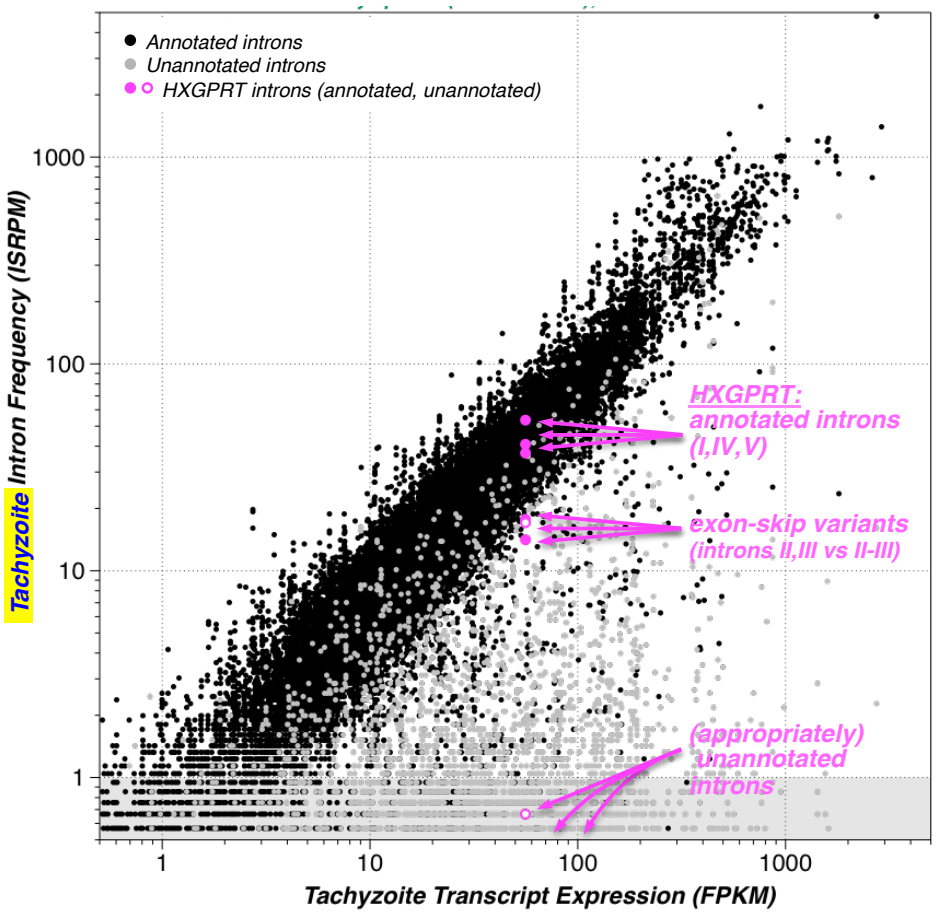


Figure 2.4. The abundance of annotated introns is overwhelmingly correlated with total transcript abundance.

Intron-spanning read abundance (ISRPM) as a function of transcript coverage (FPKM), for annotated introns (*black*) and novel (unannotated) introns (*gray*), in *T. gondii* tachyzoites. Just 31K introns are reproducibly observed and map to annotated genes at the expected abundance levels. Most of the introns observed at levels comparable to transcript abundance are annotated, *i.e. black dots* lie along the diagonal. *Magenta dots* highlight HXGPRT introns (see Fig 2.1), including (from top to bottom) the invariably-spliced introns IV & V and the slightly less-efficiently excised 5'UTR intron I; alternatively-spliced introns II, III and the intron II-III exon skip isoform; and several low abundance unannotated introns (mostly off-scale).

Assuming the reference annotation provides some approximation of the truth (even if it cannot be considered a gold standard), and that systematic errors in gene model

prediction are not overwhelming, the current annotation can be used to define parameters for intron abundance and splicing efficiency that minimize false discovery (Supplementary Fig S1). As quantified in Table 2A, the majority of introns that are efficiently excised in tachyzoite-stage parasites (ISRPM/FPKM ≥0.2) are annotated (98%), and the majority of annotated introns (81%) are efficiently excised. Conversely, the majority of low abundance introns (ISRPM <1, regardless of FPKM) are not included in the reference annotation (57%), and the majority of unannotated introns are low abundance (88%), even after exclusion of >2.5M very low abundance or poor quality unannotated introns (see above).

Table 2A. Intron abundance & annotation in various samples

Intron Annotation	Abundance (ISRPM)	Excision eff'y (ISRPM/FPKM)	Tachyzoites (strain CZ-H3)	Gametocytes (CZ-H3, d7)	Max from ANY sample
Annotated 39,653	High (≥10)	OK (≥0.2)	18,715	15,592	29,584
	Med (1-10)	"	12,805	18,259	7,325
	Low (0.1-1)	"	627	1,010	51
	High (≥10)	Low (<0.2)	51	37	654
	Med (1-10)	"	370	245	413
	Low (0.1-1)	"	197	339	141
	Very Low (<0.1)	NA	6,888	4,171	1,485
Unannotated 11,834	High (≥10)	OK (≥0.2)	154	112	1,108
	Med (1-10)	"	306	706	1,739
	Low (0.1-1)	"	44	163	53
	High (≥10)	Low (<0.2)	72	102	537
	Med (1-10)	"	900	441	3,538
	Low (0.1-1)	"	881	1,295	1,301
	Very Low (<0.1)	NA	9,477	9,015	3,558
Intergenic or ambiguous 8,268	High (≥10)		75	210	1,230
	Med (1-10)		526	1,442	3,809
	Low (0.1-1)		392	1,193	1,005
	Very Low (<0.1)		7,275	5,423	2,224
		Total:	59,755	59,755	59,755
Extremely Low (<<0.1)		Previously excluded (unannotated):			2,535,138

Table 2B. Alternative splicing

Reference Intron			Alternative Intron (ISRPM ≥1 only)						
				Unannotated	Annotated				
	Abundance (ISRPM)	Excision eff'y (ISRPM/FPKM)	None	Low Eff'y (<0.2)	High Eff'y (ISRPM/FPKM ≥ 0.2) Ref > Alt	Alt > Ref	Low Eff'y (<0.2)	Ref > Alt	Alt > Ref
Annotated 39,653	High (≥10)	OK (≥0.2)	23,983	4,569	838	194			
	Med (1-10)	"	6,957	72	164	132			
	Low (0.1-1)	"	47	2	0	2			
	High (≥10)	Low (<0.2)	471	123		60			
	Med (1-10)	"	271	37		105			
	Low (0.1-1)	"	69	13		59			
	Very Low (<0.1)	NA	1,079	56		350			
Unannotated 11,834	High (≥10)	OK (≥0.2)	387	67	56	48	114	151	285
	Med (1-10)	"	754	31	35	65	50	123	681
	Low (0.1-1)	"	28	0	1	4	3	1	16
	High (≥10)	Low (<0.2)	95	67		41	38		296
	Med (1-10)	"	2,829	313		190	84		122
	Low (0.1-1)	"	946	59		96	39		161
	Very Low (<0.1)	NA	634	123		185	68		2,548
Intergenic or ambiguous 8,268	High (≥10)		1,220		5	5			
	Med (1-10)		3,771		6	32			
	Low (0.1-1)		989		0	16			
	Very Low (<0.1)		2,209		0	15			
									Total: 59,755
Extremely Low (<<0.1)			2,535,138	Previously excluded (unannotated)					

Table 2. Intron abundance and alternative splicing in *T. gondii*.

Red and *orange* lettering indicates likely false-positives, *i.e.* annotated introns never observed and/or excised in any sample in this study; *green* lettering indicates false negatives, *i.e.* unannotated introns observed at levels consistent with transcript abundance. **A**, Intron abundance in tachyzoites, gametocytes, or any sample, stratified by splicing efficiency and annotation status. *Green, red* and *blue* shading reflects the relative abundance of annotated, unannotated, and non-genic introns in each sample. **B**, Stratification of all introns (from right-most column in panel A) based on the abundance, splicing efficiency and annotation status of alternative (overlapping)

introns. *Orange* shading reflects unannotated introns that should probably be considered alternative splice junctions; *green,* currently unannotated introns that should replace the existing annotation; *red,* annotated introns that should be removed from the current annotation; *blue,* annotated introns that should be downgraded due to more abundant unannotated alternatives.

All of the HXGPRT introns discussed above (*cf.* Fig 2A) are appropriately represented in Fig 2.4 (*magenta circles;* filled = annotated; open = unannotated). The three introns that are always excised (I, IV & V) lie close to the ISRPM = FPKM diagonal, while introns II & III, which define the exon cassette responsible for HXGPRT isoform II, are less abundant (ISRPM/FPKM ~0.3), as is intron II-III, representing the alternative exon-skip variant responsible for HXGPRT isoform I (open circle). All three of these alternatively-spliced introns still lie within the true-positive sector (ISRPM >1 and >20% FPKM), but less plausible HXGPRT introns do not (including several that are off-scale on this plot). While most annotated introns lie within the true-positive sector in this CZ-H3 tachyzoite sample, 8133 appear as possible false positives in Fig 2.4, *i.e.* ISRPM <1 and/or ISRPM/FPKM <0.2. Considering other samples, however, reveals that the majority of these introns are expressed in other life cycle stages, as discussed below.

For example, while a similar distribution of annotated and unannotated introns is observed in gametocyte stage parasites (Fig 2.5A, Table 2A), most of the apparent tachyzoite stage false-positives are abundantly expressed and efficiently excised in gametocytes (*orange* datapoints in Fig 2.5). ~79% of all annotated introns are supported in tachyzoites, and 85% in gametocytes, but only 73% were observed in both life cycle stages, while 6-12% were stage-specific (Table 2A). Combining all 20 of the high quality datasets used in this study (green shading in Table 1) and plotting the maximum evidence for each intron (in any experiment) against transcript abundance in that experi-

ment (Fig 2.5B), provides strong support for 93% of all annotated introns at ISRPM ≥1 and ISRPM/FPKM ≥0.2, and 75% at ISRPM ≥10 (Table 2A), *i.e.* there are relatively few false positives in the current *T. gondii* reference annotation. It is likely that more extensive sequencing, including additional strains, life cycle stages, conditions and treatments would reduce this number still further.

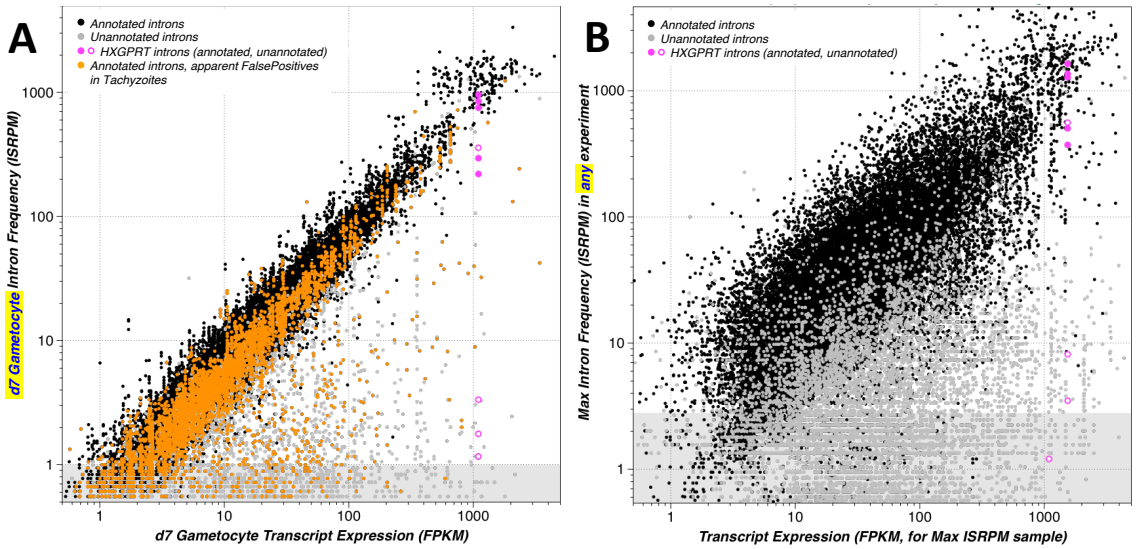


Figure 2.5. Genome-wide identification of implausible annotated introns (FP). Intron-spanning read abundance (ISRPM) as a function of transcript coverage (FPKM) for annotated introns (*black*) and novel (unannotated) introns (*gray*). **A**, *T. gondii* gametocytes display a similar pattern with most introns observed at levels comparable to transcript abundance being annotated, *i.e. black dots* lie along the diagonal, but many introns not expressed in tachyzoites (black dots at bottom or below the diagonal in Fig 2.4) are expressed in gametocytes (*orange*), often at transcriptionally-appropriate levels. **B**, ISRPM *vs* FPKM values for all putative introns, in whichever sample that intron is maximally evident; note that this provides further separation of black and gray dots, leaving few false positives.

Potential false negatives, *i.e.* unannotated introns that map to genes with a maximum abundance consistent with mRNA levels (~7% of plausible intron candidates), are represented by *gray* datapoints overlapping the dominant black cloud in Fig 2.5B. It is

interesting that most of these candidate splice junctions display somewhat lower abundance and/or inefficient excision relative to expectations based on FPKM values, *i.e.* they are slightly offset towards lower edge of the black diagonal cloud representing annotated introns (true positives). This observation suggests that they may represent subdominant splice variants rather than fully false negatives, as is the case for introns attributable to HXGPRT exon-skip variants. (Note that HXGPRT is very highly expressed in gametocytes, at levels ~20-fold greater than in tachyzoites; compare pink dots in Fig 2.5 & 2.4.)

To directly assess the relationship between possible false negatives and existing intron annotation, we compared the maximal excision efficiency (ISRPM/FPKM) for each intron (vertical axis in Fig 2.6) with the excision efficiency of the maximum *alternative* overlapping intron (horizontal axis). The four quadrants of this figure, defined by excision efficiencies of greater or less than 20% in the reference *vs* overlapping alternative introns represent:

- *Top Left:* Introns expressed and excised at comparable levels (ISRPM/FPKM ≈ 1) and overlapping alternative introns that are ~5-5000-fold less efficiently spliced. This quadrant is dominated by annotated introns (*black*), including the efficiently-spliced HXGPRT introns I, IV & V. Many additional annotated introns lack any plausible overlapping alternative (off-scale to the left).
- *Bottom Right:* Introns that are inefficiently spliced but overlap alternatives excised at expected levels, *i.e.* the reciprocal of the above. This quadrant is dominated by unannotated introns (*blue*), including implausible alternatives to HXGPRT introns I, IV & V (*cf.* Figs 1A&C), and many more off-scale low.
- *Bottom Left (gray shading):* Low abundance and/or inefficiently spliced introns that overlap alternatives that are also low abundance and/or inefficiently spliced (*gray*).

- *Top Right (yellow shading):* Introns that are abundant and efficiently spliced, and overlap alternatives that are also abundant and efficiently spliced, such as those for HXGPRT introns II & III (*solid magenta dots*) and the exon-skip variant represented by HXGPRT intron II-III (*open magenta circle*). (Note that stage-specific alternative exons would be expected to be even more efficiently excised.)

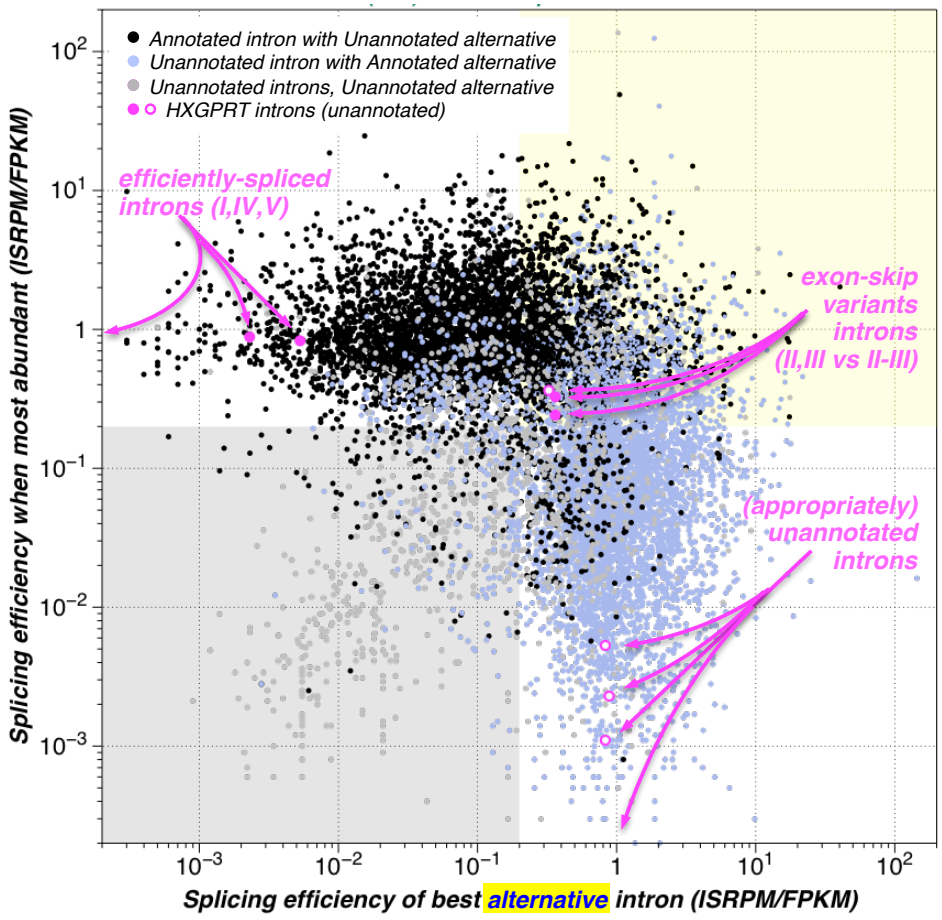


Figure 2.6. Genome-wide identification of likely unannotated introns (FN), including alternatively-spliced isoforms.

Plotting splicing efficiency (ISRPM/FPKM) from Fig 2.5B against splicing efficiency for the most prevalent overlapping alternative (Additional File 1) identifies unannotated introns that should be considered for inclusion in the reference annotation (~7.2% of annotated introns), including likely splice variants (~3.6% of annotated introns; *yellow quadrant*); *Black dots* indicate annotated introns overlapping unannotated alternatives; *blue dots* indicate unannotated introns overlapping

annotated alternatives, and *gray dots* denote unannotated introns overlapping unannotated alternatives. See text for further discussion.

Summary statistics for overlapping introns are provided in Table 2B. Considering Tables 2A & 2B in aggregate, and comparing with Fig 2.6, permits an objective analysis of current annotation status, using metrics that should also be applicable to other systems:

- *True Positives:* 93% of *T. gondii* introns annotated in ToxoDB release 28 are abundantly or moderately expressed and efficiently excised in at least one experimental sample (*black* datapoints in the upper half of Fig 2.6): 29,584 display maximal abundance ≥10 ISRPM and splicing efficiency (ISRPM/FPKM) ≥20%; 7325 display maximal ISRPM from 1-10 (and ISRPM/FPKM ≥20%). 84% of these (23,983 + 6957) show no overlap with any plausible alternative intron (off-scale left in Fig 2.6). 5969 overlap plausible alternatives, but 78% of those alternatives (4569 + 72) are only weakly expressed and/or inefficiently excised (horizontal black cloud in Fig 2.6). Just 1328 plausibly-expressed and efficiently-spliced annotated introns were found to overlap plausible unannotated alternatives (black dots within the yellow shaded region). Among these, the annotated intron was most prominent 75% of the time (838 + 164; orange shading in Table 2B; above the diagonal in Fig 2.6). Only 326 (194 + 132) overlap unannotated introns that are more prominent than well-supported annotated introns (green shading in Table 2B; black dots below the diagonal in the yellow-shaded region of Fig 2.6).

- *False Positives:* 2744 annotated introns (*red & orange lettering* in Table 2A) were identified as likely false positives due to low evidence (1677) and/or inefficient excision (1208), and are represented by black dots in the lower half of Fig 2.6. 576 overlap plausible unannotated introns that could be considered as replacement annotation (red/orange letting and green shading in Table 2B; black dots in the lower right-hand quadrant of Fig 2.6). Many of these share either a donor or acceptor splice site with the unsupported annotation (see below).

- *False Negatives:* 6922 unannotated introns map to genes in the current reference annotation and are abundantly expressed in at least one sample, but only 41% of these (1108 + 1739) are also efficiently excised. Most (1444; 51%) overlap plausible alternative introns (gray & blue dots in the upper half of Fig 2.6), which are usually more probable annotated introns (681 + 285 = 966; blue dots below the diagonal in the yellow shaded region of Fig 2.6). A few (151 + 123 = 274) represent introns that are more probable alternatives to the existing annotation (blue dots above the diagonal in the yellow shaded region in Fig 2.6; blue shading in Table 2B), and fewer still should probably supplant the existing annotation (just 114 + 50 = 164; blue dots in the upper-left quadrant, the reciprocal of black dots in the lower right-hand quadrant, described above; red shading in Table 2B). In addition, 302 plausible unannotated introns overlap unannotated alternatives (gray dots in the upper half of Fig 2.6), and 1141 high confidence unannotated introns do not overlap other plausible introns (Table 2B; off-scale left in Fig 2.6).

- *True Negatives:* Even excluding the millions of very low abundance junctions previously discarded, most unannotated introns are not significantly expressed and/or processed, in any sample (black numbers in the right-hand column of Table 2A).

- *Intergenic Introns:* Introns observed in RNA-seq experiments that fail to map to annotated genes are relatively rare, but 1230 + 3809 = 5039 display plausible abundance (Table 2B), suggesting extended gene models or previously unannotated genes. Many of these are most evident in non-tachyzoite stage parasites, suggesting stage-specific expression, and reflecting the dominance of data from the acutely-lytic *in vitro* tachyzoites in generating the current reference annotation.

Implications for *T. gondii* annotation

Exploiting the parameters defined above, various transformations and filtering of data in Additional File 1 were used to identify potential cases of erroneous annotation in the reference *T. gondii* genome, including alternative splice variants and novel gene models. Representative examples are provided in Figs 2.7-11 (and highlighted in Figs

2.12 & 13). These results are currently being incorporated into the reference genome database.

*Tg*ME49_234450 (Fig 2.7) is annotated as ribosomal small subunit protein Rps15A, based on highly conserved protein sequence motifs across all species. The current structural annotation describes two introns, but overlapping alternatives (defined based on the analysis presented in Fig 2.6) reveal a skipped exon within the first intron (*magenta star*) in ~30% of tachyzoite transcripts. This organization (and abundance) is similar to the HXGPRT cassette exon (Fig 2.1), except that the reference annotation for HXGPRT includes the cassette exon, whereas the annotated RPS15A transcript does not. In contrast to the 147 nt HXGPRT cassette, which encodes 49 amino acids, the 69 nt RPS15 cassette exon is predicted to result in premature protein translation, due to an internal stop codon (Supplementary Fig S2). Close inspection of the genome sequence and multiple sequence alignments reveals that the annotated translation initiation site is likely incorrect. An alternative ATG immediately downstream of the cassette exon offers better sequence context for translation and protein stability (Matrajt et al. 2004), and yields a product that aligns with RPS15A proteins from other species across its entire 130 amino acid length (Fig 2.7, bottom).

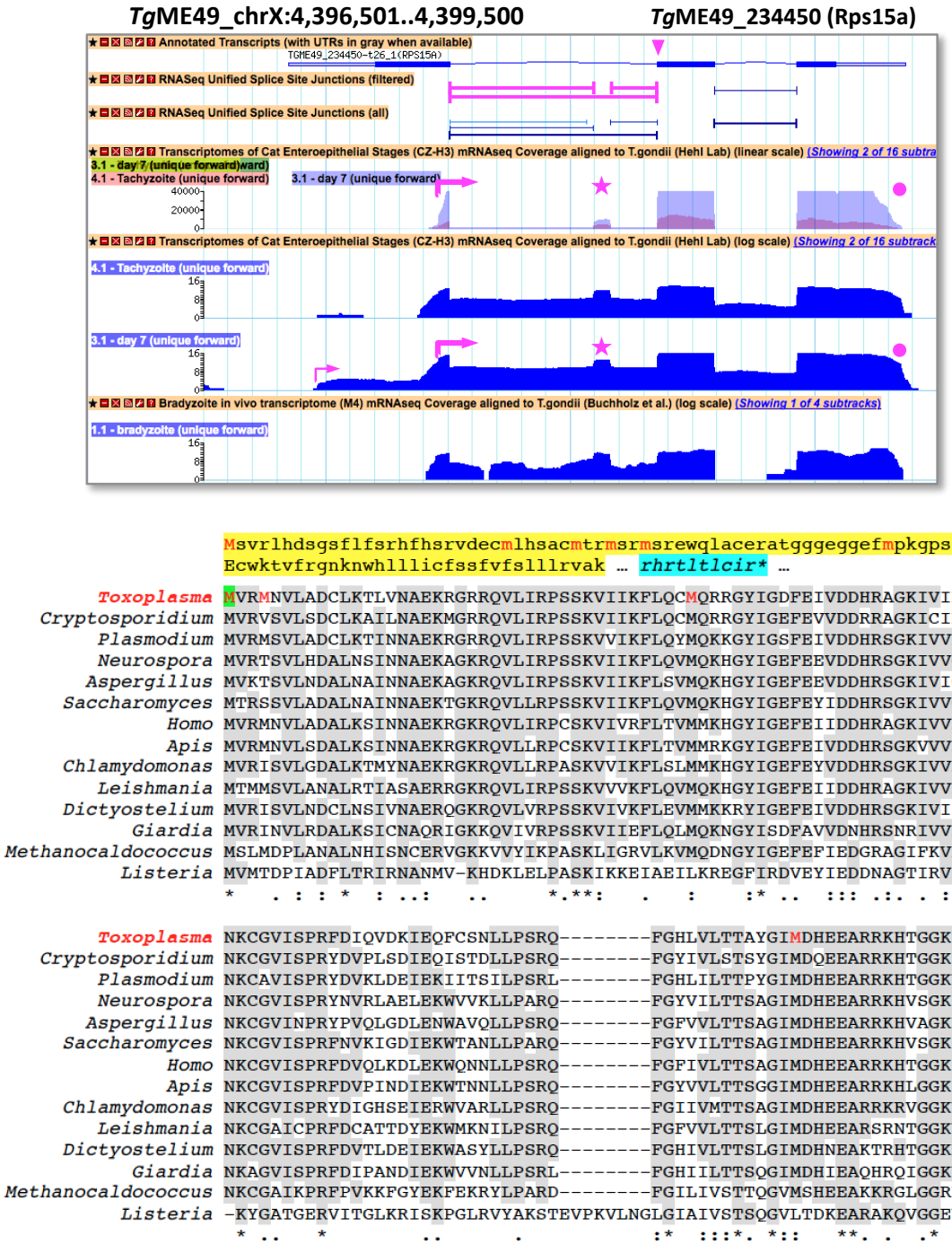


Figure 2.7. Genome browser view & multiple sequence alignment of *Tg*Rps15a.

Top, Genome annotation and selected RNA-seq tracks from tachyzoites, bradyzoites, and/or gametocytes (as labeled). Note that solid blue tracks are scaled logarithmically while tracks

shown in semi-transparent overlay are linear (also note different scales, selected to highlight alternative splicing). As in Fig 2.1, *arrows* and *circles* indicate transcript termini, *stars* indicate variants discussed in the text. In contrast to Fig 2.1, two tracks display candidate introns (brackets): the 'filtered' track is restricted to display only those introns that pass the abundance and efficiency criteria described in this dissertation. ***Bottom,*** Multiple sequence alignment of *Tg*RPS15a with various other species. Current annotation includes a 101 amino acid N-terminal exon (*yellow*) upstream of the most plausible initiation codon (*green*); inclusion of the alternatively-spliced cassette exon noted above would introduce 10 amino acids ending in a premature termination codon (*turquoise*).

Moreover, while proteomics evidence provides ample support for the body of the protein (Wastling et al. 2009; and other data in ToxoDB.org), there is no evidence for translation of the 101 amino acid N-terminal extension predicted by the current annotation. In sum, there is strong support for an exon-skip variant at this locus, and inclusion of the cassette may be stage-specific (more common in bradyzoites, less common in gametocytes), but these alternative splice products likely affect the 5'UTR only. As indicated by the RNA-seq data, and supported by the presence of chromatin activation marks (Gissot et al. 2007; ToxoDB.org) the annotated transcription start site corresponds to the longest potential transcript, but not the most abundant. The annotated UTR represents <0.1% of transcripts; most transcription initiates >500 nt downstream, just upstream of intron I.

The availability of RNA-seq data dramatically improves *T. gondii* UTR annotation (*cf.* Fig 2.2), but in general, the algorithms employed (Bernal, Crammer, and Pereira 2012; Bernal and Pereira 2012) appear to be too greedy, annotating the longest, rather than the most abundant UTRs. Transcript initiation sites, including the use of stage-specific promoters, can dramatically impact gene models, as it is of course impossible to excise sequences that are not transcribed. Transforming the reference data (Additional

File 1) to examine stage-specific intron usage (see below, Fig 2.13) highlights Isocitrate Dehydrogenase (*Tg*ME49_266760), among many other genes. As shown in Fig 2.8, this gene appears to be transcribed from two promoters: a low-level constitutive promoter, and a 20-fold more potent gametocyte-specific promoter ~800 nt downstream, within intron I.

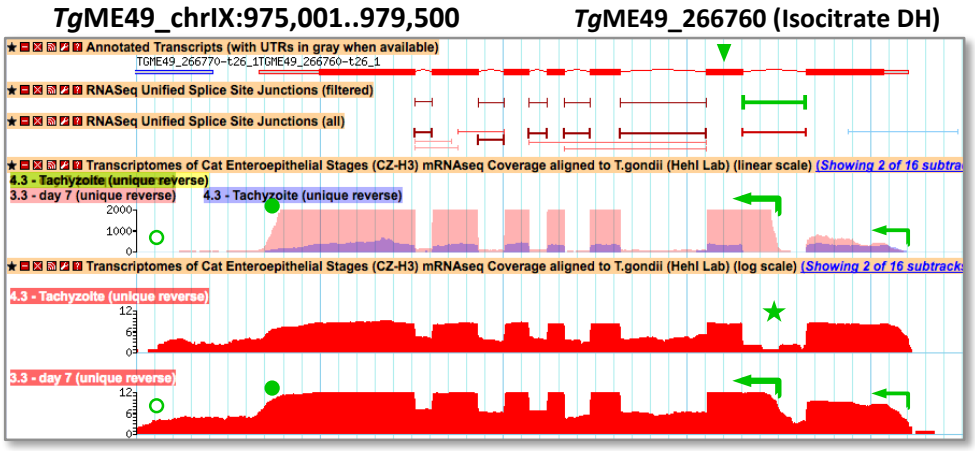


Figure 2.8. Genome browser view of alternatively-spliced Isocitrate DH.
Genome annotation and selected RNA-seq tracks from tachyzoites and gametocytes. Note that red tracks are scaled logarithmically, while tracks shown in semi-transparent overlay are linear (also note different scales, selected to highlight alternative splicing). *Arrows* and *circles* indicate transcript termini, *stars* indicate variants discussed in the text. Two tracks display candidate introns (brackets): the 'filtered' track is restricted to display only those introns that pass the abundance and efficiency criteria described in this dissertation.

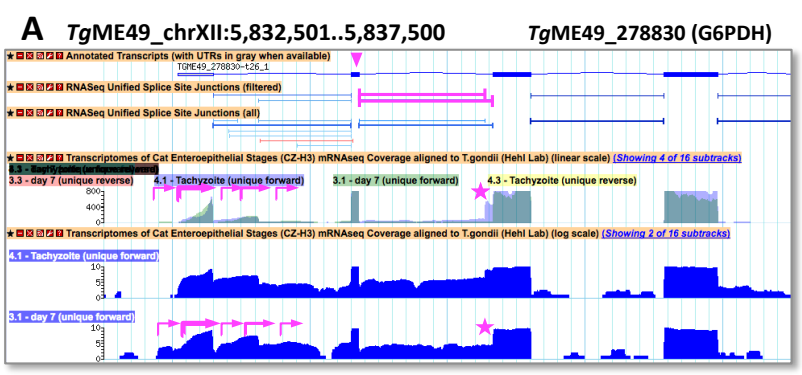
As a result, tachyzoite transcripts match the annotated gene model, but gametocyte transcripts lack exon I, initiating within the intron and therefore containing a longer "second" exon. As observed for RPS15A (Fig 2.7), multiple sequence alignment (Supplementary Fig S3) and analysis of available proteomics (Xia et al. 2008) data suggests that the annotated translation initiation codon for *Tg*ME49_266760 is incorrect, and

translation begins within the second exon, downstream of intron I, yielding the same mature protein product in both tachyzoite and gametocyte-stage parasites.

Although relatively infrequent, alternative splicing that affects the protein coding sequence is occasionally observed, as in *Tg*ME49_278830, which encodes glucose-6-phosphate dihydrogenase (Fig 2.9A, displaying just the first 4 exons of this 21 exon transcript). Multiple promoters yield a diversity of 5' ends, affecting the selection of splice donors for intron I (which lies entirely within the UTR). Two alternatives were also observed for Intron II, which lies within the protein coding sequence: the annotated intron predominates, but a splice acceptor variant is also observed, resulting in 60 nt of additional coding sequence adding 20 amino acids to the mature protein (*magenta star*) (Supplementary Fig S4). In contrast to the examples cited above, the annotated translation start displays plausible sequence context (although an in-frame ATG 48 nt downstream within the same exon is even better), and no alternative initiation codon is present upstream of highly conserved protein sequences. Experimental proteomics evidence also confirms supports usage of the second ATG (Xia et al. 2008).

As indicated in Fig 2.9B, the alternative (longer) G6PDH transcript was observed in tachyzoites from multiple parasite strains, but not in mature bradyzoites, gametocytes, or unsporulated or sporulated oocysts (also seen in bradyzoite samples generated *in vitro*, which contain numerous tachyzoites; Dzierszinski et al. 2004; Soete, Camus, and Dubremetz 1993). Interestingly, in two separate time courses examining transcription at various times post-infection, utilization of the alternative splice acceptor was most common early during intracellular replication (see Supplementary Fig S5 data for all

RNA-seq data on this differentially-spliced intron, including multiple strains with lower coverage data).



B

TgME49_278830 (G6PDH) intron 2

Samples	#	Annotated	Alternative	Ratio
All Tachyzoites	51	1062	646	0.61
RH/GT1	9	150	80	0.53
ME49/CZH3	9	290	294	1.01
VEG	10	425	213	0.50
hr 2	2	15	32	2.13
hr 4	2	15	25	1.67
hr 8	2	4	8	2.00
hr 16	2	12	12	1.00
hr 36	2	160	92	0.58
hr 44	2	123	87	0.71
All Bradyzoites	3	175	63	0.36
In vitro Bz	1	79	58	0.73
In vivo Bz	2	96	5	0.05
All Gametocytes	3	548	4	0.01
d3	1	21	1	0.05
d5	1	217	2	0.01
d7	1	310	1	0.00
All Oocysts	3	900	13	0.01
d0	1	11	0	0.00
d4	1	362	1	0.00
d10	1	527	12	0.02

Figure 2.9. Genome browser view of alternatively-spliced G6PDH and Intron II abundance. **A,** Genome annotation and selected RNA-seq tracks from tachyzoites and gametocytes. Note that blue tracks are scaled logarithmically, while tracks shown in semi-transparent overlay are linear (note different scales, selected to highlight alternative splicing). *Arrows* and *circles* indicate transcript termini, *stars* indicate variants discussed in the text. Two tracks display candidate introns (brackets): the 'filtered' track is restricted to display only those introns passing the abundance and efficiency criteria described in this thesis. **B,** G6PDH Intron II abundance. Intron excision frequency for G6PDH splice acceptor variants for all RNA-seq samples (Table 1), stratified by parasite strain, time post-infection (for tachyzoites), and life cycle stage.

As noted in Table 2, several thousand introns map to the *T. gondii* genome outside of annotated genes, and display abundance consistent with splice sites for known genes. In order to estimate the number of unannotated protein-coding genes in the parasite genome, we examined a randomly selected subset of all intergenic introns, and also a subset of those with ISRPM >35. In addition, we manually curated all intergenic regions within a 1 Mb chromosomal span. All of these analyses suggest that the current reference genome annotation is lacking ~300 protein-coding genes (including single exon genes), most of which are specifically expressed in non-tachyzoite stages, for

which data has only recently become available (from this study and others). These unannotated genes account for ~1/3 of all plausible intergenic introns. For example, Fig 2.10 shows a 5-exon gene on chromosome Ia that is expressed in gametocytes only (*green shading*). An additional third of intergenic introns can be attributed to unannotated exons extending existing gene models (often as UTR exons), and most of the remaining third are associated with non-coding RNAs of unknown significance (and often inefficiently excised).

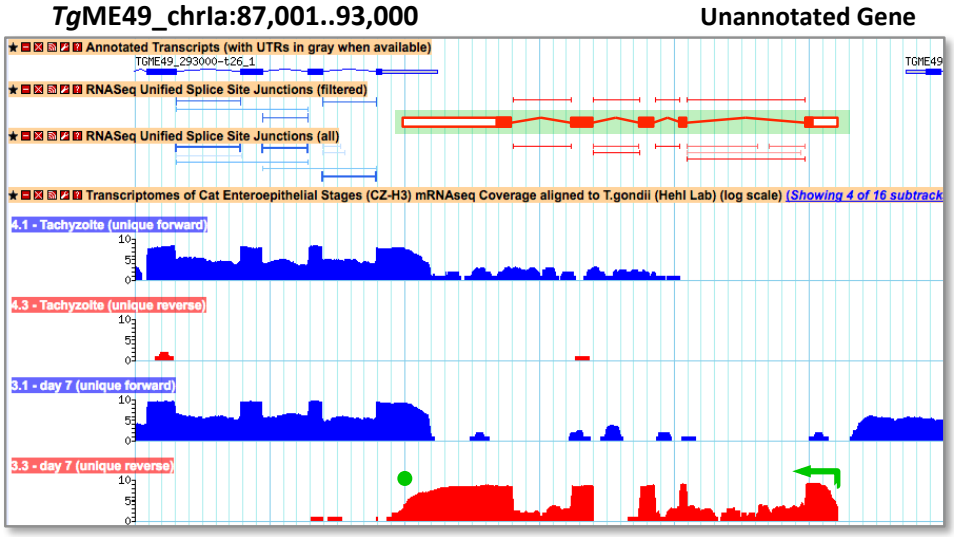


Figure 2.10. Genome browser view of an unannotated gene.
Genome annotation and selected RNA-seq tracks from tachyzoites and gametocytes. Note that blue and red tracks are scaled logarithmically. *Arrows* and *circles* indicate transcript termini. Two tracks display candidate introns (brackets): the 'filtered' track is restricted to display only those introns that pass the abundance and efficiency criteria described in this thesis.

Overall, the most common errors associated with existing annotation derive from inaccurate annotation of transcript termini, which are often stage-specific. As noted above, this may impact intron usage, splicing efficiency, and the selection of translation

start sites. In the absence of functional evidence, however, it is not necessarily clear that alternative starts and stops are biologically relevant, or should be annotated. For example, the dynein heavy-chain family protein *Tg*ME49_309980 (Fig 2.11A) is a 63 exon gene, spanning nearly 50 kb. Despite low expression levels (FPKM <10), all 62 introns appear to be accurately annotated, but tachyzoite stage-specific termini suggest a much shorter transcript in this stage, including exons XXXVII-LIII only (or perhaps XIX-LXV), and lacking many domains likely to be essential for function. From the available data (all based on steady-state RNA, as ribosomal profiling data is not yet available for *T. gondii*), it is unclear whether the different transcript termini inferred from tachyzoite *vs* gametocyte RNA-seq data reflect differences in transcriptional initiation and termination, or differences in mRNA stability and degradation, but chromatin marks provide some evidence in support of a shorter tachyzoite-specific transcript (not shown). In addition, a tachzyoite-specific antisense RNA transcript (lncRNA; green box in Fig. 2.11) could potentially play a role in regulating mRNA stability (discussed further below).

While most genes in the *T. gondii* genome are accurately annotated, and most high abundance unannotated introns are readily interpretable (as noted above), some genes will remain refractory to annotation without additional experimental data. For example, *Tg*ME49_270520 (Fig 2.11B) encodes a 'hypothetical protein' of no known function that is most abundant in gametocytes but expressed in all life cycle stages except unsporulated oocysts. The current reference annotation includes 6 exons, but at least 2 additional exons are evident (*green stars*); analysis of intron-spanning reads suggests that exons IV-VII may be included in the mature mRNA in various permutations and combinations, possibly using different combinatorial patterns in different life cycle stages.

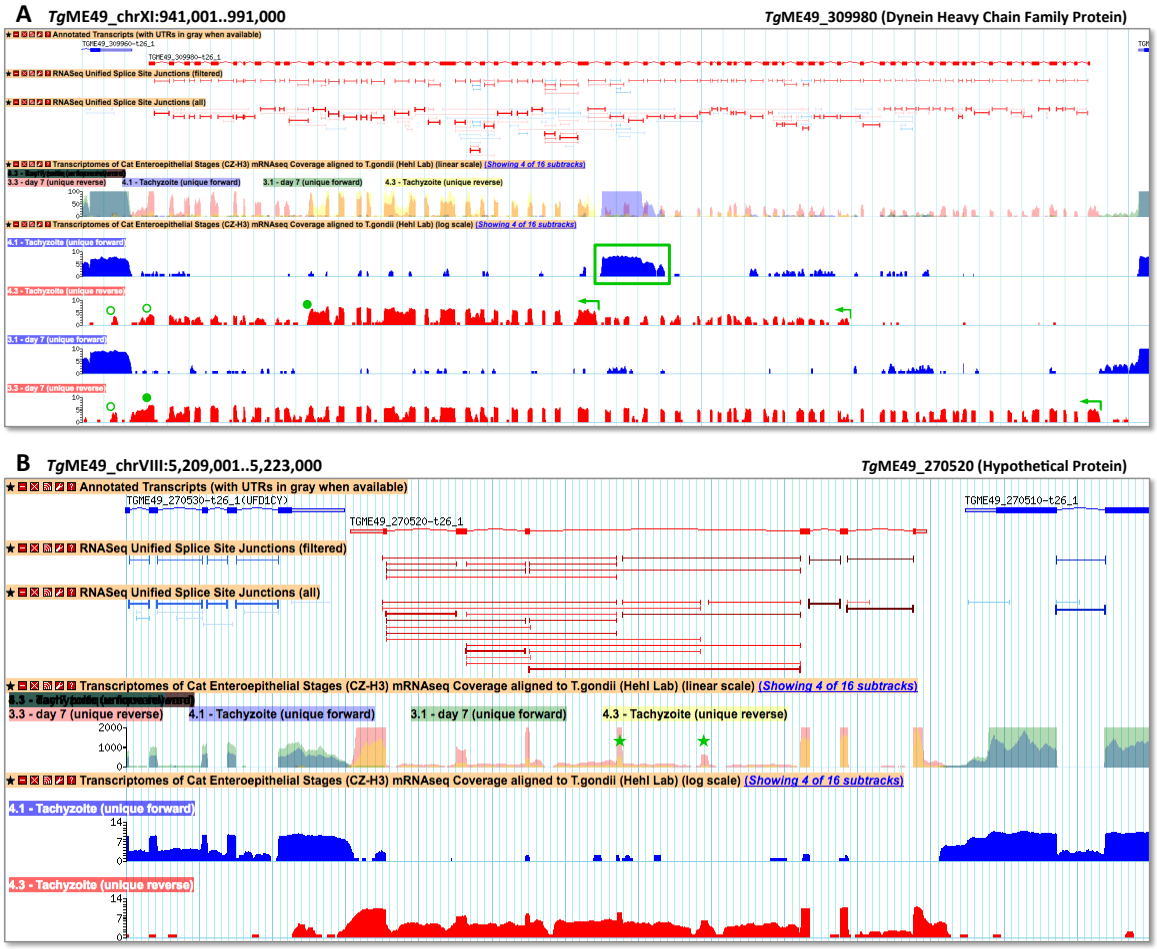


Figure 2.11. Genome browser view of alternatively-spliced transcripts that are difficult to confidently annotate.

Genome annotation and selected RNA-seq tracks from tachyzoites and gametocytes. Note that blue and red tracks are scaled logarithmically, while tracks shown in semi-transparent overlay are linear (also note different scales, selected to highlight alternative splicing). *Arrows* & *circles* indicate transcript termini, *stars* indicate variants discussed in the text. Two tracks display candidate introns (brackets): the 'filtered' track is restricted to display only those introns that pass the abundance and efficiency criteria described in this dissertation.

Figure 2.12 reproduces Fig 2.6, highlighting the expression of annotated and unannotated introns for some of the alternatively-spliced genes discussed above. Note that most annotated introns (*closed circles*) only overlap alternatives at implausibly low

abundance, *i.e.* they map to the upper left quadrant, while most unannotated introns (*open circles*) map to the lower right quadrant, *i.e.* they overlap annotated introns that are much more efficiently expressed and spliced.

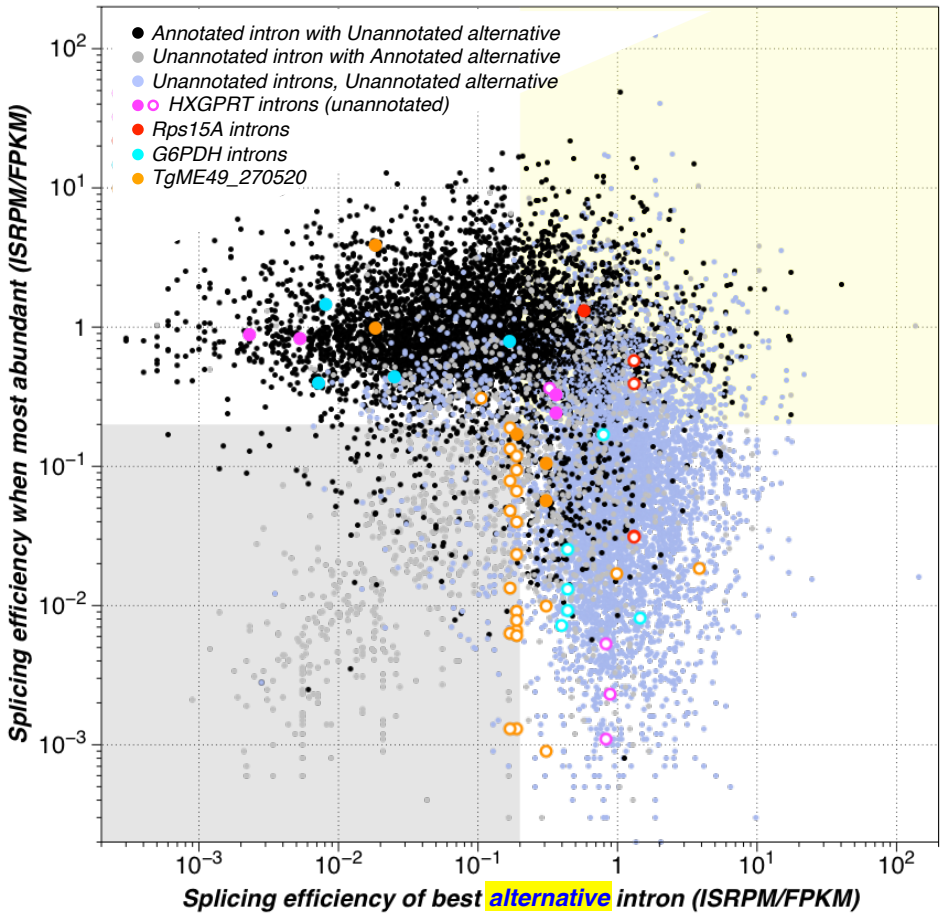


Figure 2.12. Identification of alternative-spliced introns.
Maximum splicing efficiency (ISRPM/FPKM) for each putative intron *vs* its most prominent overlapping alternative intron (replica of Fig 2.6). *Black,* annotated introns overlapping unannotated alternatives; *Blue-gray,* unannotated introns overlapping annotated alternatives; *Gray,* unannotated introns overlapping unannotated alternatives. Colored dots highlight introns associated with *Tg*ME49_200320 (HXGPRT; *Magenta* … see Fig 2.1), *Tg*ME49_234450 (Rps15a; *Red* … see Fig 2.7), *Tg*ME49_278830 (G6PDH; *Turquoise* … see Fig 2.9), *Tg*ME49_270520 (hypothetical protein; *Orange* … see Fig 2.11B).

Alternatively-spliced introns map to the *yellow shaded area*, including the two annotated HXGPRT introns and the single unannotated cassette exon-spanning intron (*magenta*), the single annotated exon-spanning intron in Rps15A and the two unannotated introns flanking the cassette exon (*red*), and both the annotated and unannotated splice acceptor variants for G6PDH (*turquoise*). Additional annotated HXGPRT and G6PDH introns are efficiently spliced, but lack any alternatives, and are therefore off-scale to the left. Several unannotated G6PDH introns overlap the first annotated intron, within the 5'UTR (see Fig 2.9), and map to the lower right quadrant. The first two *Tg*ME49_ 270520 introns are excised in all splice variants (see Fig 2.11B) and are represented by *solid orange dots* in the upper left quadrant, but other annotated and many unannotated introns combining cassette exons in various ways yield a diversity of introns displaying low apparent splicing efficiency.

Figure 2.13 highlights stage-specific alternatively spliced introns, by comparing the intron excision efficiency (ISRPM/FPKM) of tachyzoites *vs* gametocytes. For example, Isocitrate DH intron I (*Tg*ME49_266760; left-most *green* dot) is inefficiently excised in gametocytes only because the gametocyte stage-specific promoter lies within the first intron (see Fig 2.8). Many such intron isoforms map to UTRs, often incorrectly annotated (as in this case). Because only the central portion of *Tg*ME49_309980 (Dynein HC; see Fig 2.11A) is transcribed in tachyzoites, gene-level FPKM values are low, central introns appear to be excised with efficiencies >1, and introns outside this region appear to be excised unusually poorly (*blue* dots in Fig 2.13). The unannotated intron acceptor variant in G6PDH (*turquoise;* see Fig 2.9A) is not observed in gametocytes (Figure 2.9B), and therefore lies on the vertical axis in Fig 2.13.

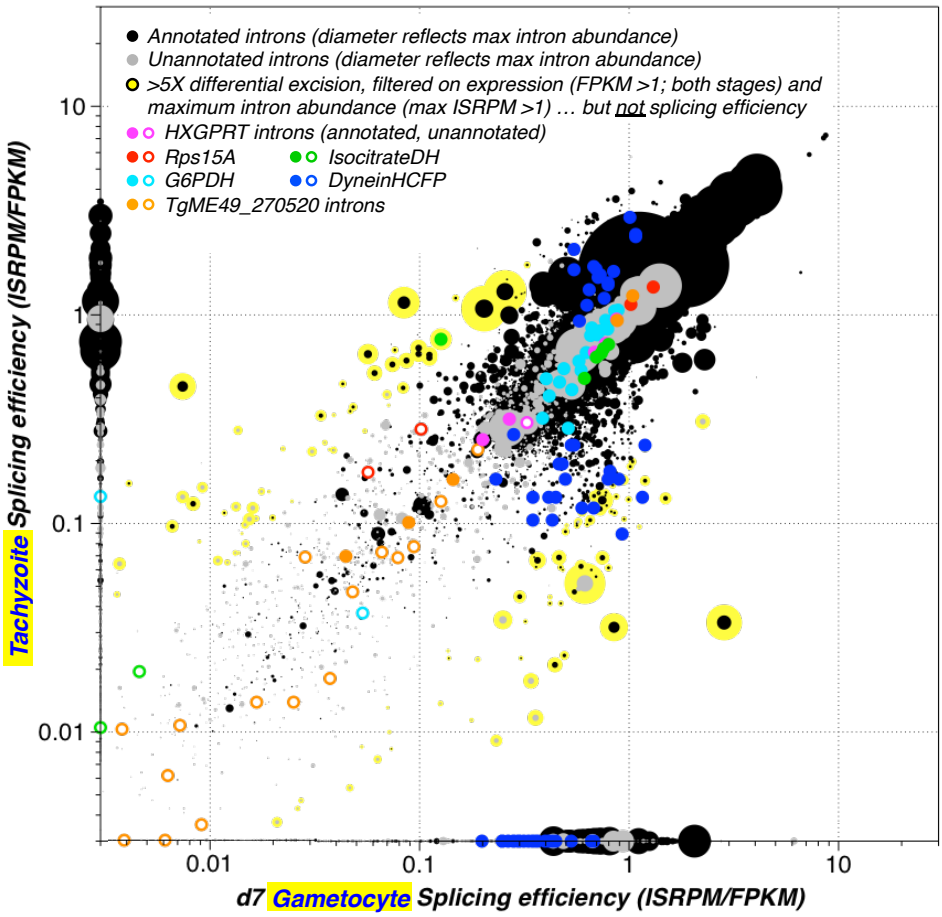


Figure 2.13. Identification of stage-specific alternative-spliced introns.
The size of the dots in this panel reflects the maximal ISRPM observed in these samples. *Yellow* highlights introns displaying FPKM values >1 in both samples, maximum intron abundance (in either sample) >1, and >5-fold differential intron excision efficiency between tachyzoites and gametocytes. Introns labelled in magenta, red, turquoise and orange show no significant stage-specificity in intron excision, except as noted in the text.

Discussion

A previous report has demonstrated the value of RNA-seq data for improving *T. gondii* genome annotation, including the recognition of alternatively-spliced isoforms (Hassan et al. 2012). Because our preliminary analysis revealed that strand-specific

RNA-seq data is substantially more valuable for gene model assessment (*cf.* Fig 2.3), detailed studies of structural annotation and alternative splicing were carried out using the 20 highest quality, strand-specific datasets (total of >500M reads, including >60M intron-spanning reads; Table 1). Results were then integrated with all available information, including non-strand-specific and/or lower depth RNA-seq data(Lorenzi et al. 2016; Minot et al. 2012; Reid et al. 2012; Pittman, Aliota, and Knoll 2014; Hehl et al. 2015), as well as microarray (Fritz et al. 2012; Grigg et al. 2001; Behnke et al. 2014; Buchholz et al. 2011; Bahl et al. 2010), SAGE tag & EST studies (Radke et al. 2005), proteomics data (Xia et al. 2008), chromatin marks (Gissot et al. 2007), population-level sequence variation (SNPs; Lorenzi et al. 2016), *etc.* These results were consistent with all previous reports on transcript variation, including alternative splicing of the HXGPRT locus (Fig 2.1; Donald et al. 1996; Chaudhary et al. 2005), and examples highlighted by Hassan *et al.* 2012.

Overall, the accuracy of existing reference annotation is quite high: >93% of all annotated introns are supported by the pooled mRNA-seq datasets. In aggregate, these samples define >2.7M intron junctions, but while almost all correspond to consensus intron boundaries and are therefore likely to be genuine, most are present at very low abundance (<<0.1 intron-spanning reads per million; ISRPM). Establishing a minimum requirement of ISRPM >1 appears to be quite conservative, removing very few annotated introns, but excluding the vast majority of unannotated introns (Figs 2.4-2.6, Additional File 1). In addition, it is also helpful to exclude introns that are inefficiently excised (low ISRPM/FPKM ratios; also known as ψ (Percent Spliced In; Venables et al. 2008), even if they are frequently observed in heavily transcribed genes. Various statistical methods have been developed to improve the estimation of ψ (Katz et al.

2010; Shen et al. 2014; Trapnell et al. 2013; Vaquero-Garcia et al. 2016), but ISRPM/FPKM serves as an adequate proxy when gene annotation is close to accurate, and is easier to automatically extract from RNA-seq mapping pipelines and genome databases. Assuming that the current *T. gondii* annotation provides a reasonable approximation of the truth, with little systematic bias, setting the minimum excision efficiency at 20% minimizes false discovery (Fig 2.6, Fig S1) … although of course some less frequent isoforms may still be functional, or may be observed at higher frequency under as-yet-untested conditions.

Plotting intron frequency *vs* transcript abundance clearly distinguishes introns that are always excised from those that are alternatively spliced, and those with low penetrance that are unlikely to yield biologically-functional transcripts (Fig 2.4; see HXGPRT as a positive control; Uhlén et al. 2015). Most of the putative false positives in this example are in fact expressed in other life cycle stages (Table 2, Fig 2.5A). Defining each intron based on the sample providing maximum support permits identification of >93% of all annotated introns (Table 2, Fig 2.5B). A modest number of annotated introns (2693) have no support and should probably be removed from the official annotation, and a similar number (2847) should be added, either as novel introns (1239), or in conjunction with other unannotated introns (204), secondarily to an annotated intron (966), as the preferred isoform (274), or in a few cases replacing an annotated intron (164). Plotting ISRPM/FPKM ratios in reference introns *vs* the most prominent alternative isoform (if one exists) provides a more sensitive means for discriminating alternatively-spliced introns (Figs 2.6, 2.12 & 2.13). These criteria are readily automated, permitting application to many incompletely annotated genomes. For example, the filtering characteristics based on those described above have recently been imple-

mented in ToxoDB build 29 (compare the two intron tracks presented in Figs 2.7-2.11 with minimally-filtered splice junction data in Figs 2.1 & 2.3), and appear applicable to *Plasmodium falciparum* as well (see Chapter 4).

Case studies reveal the same assortment of incomplete annotations observed in other eukaryotes (Mudge and Harrow 2016; Hassan et al. 2012), including UTR information, missing exons (especially at transcript termini), and the usual assortment of splice variants: alternative splice donors, splice acceptors, exon-skip variants, unexcised introns, *etc* (Figs 2.7-2.11). A plethora of overlapping but inefficiently spliced introns suggests multiple promoters, frequently associated with a UTR (*cf.* Figs 2.7, 2.8 & 2.9), and often resulting in mis-annotation of the most probable translational start (ATG). Indeed, it seems likely that improper ATG annotation constitutes the most biologically-relevant source of errors in current *T. gondii* annotation, as accurate definition of the mature N-terminus has important implications for subcellular localization predictions, and trafficking is particularly important for mediating pathogen interactions with their host cells. The sequence context of *T. gondii* initiation codons is known to be highly constrained (Matrajt et al. 2004), providing a basis for systematic improvement of CDS annotation. UTR predictions could also be improved substantially by using a less greedy algorithm to select the most abundant UTR(s), rather than the longest (*cf.* Fig 2.7). Similarly, gene annotations that incorporate the most abundant introns, rather than the one that results in the longest open reading frame, as is currently the case, would yield more accurate gene models. Additional evidence based on chromatin marks (Gissot et al. 2007) and proteomics studies (Xia et al. 2008) can provide further support for (or refutation of) candidate gene models, but ultimately, there is no substitute for biologists understanding and being able to interpret the underlying data. Without additional

experimental evidence, it is unlikely that any algorithm will be able to accurately annotate biologically significant transcriptional start sites for *Tg*ME49_288830 (Fig. 2.9), or sensible structural models for *Tg*ME49_270520 (Fig. 2.11B),

In the absence of gene assignments (and therefore FPKM values), evaluation of introns that map outside annotated genes is more difficult, but this problem could be addressed by using local analyses (Katz et al. 2010; Shen et al. 2014; Trapnell et al. 2013; Vaquero-Garcia et al. 2016). On the order of 300 protein-coding genes appear to be missing from the current annotation, but now that data is available for most life cycle stages, it is likely that many could be identified by re-running *de novo* gene predictors, informed by newly available RNA-seq datasets. This would be especially true since current *de novo* gene finders have used evidence exclusively from the tachyzoite stage and genes expressed in other non-tachyzoite stages have been predicted based on *ab initio* gene finders only. Other transcripts evident within intergenic regions (both with and without introns) may highlight long non-coding RNAs. Defining these lncRNAs has proved difficult however, as gene finders cannot rely on the same sequence features known to be useful for protein coding genes.

CHAPTER 3: MECHANISMS OF TRANSCRIPTIONAL REGULATION IN *TOXOPLASMA GONDII* (Adapted From Paper)

The development and dissemination of low-cost technologies for deep sequencing of RNAs has yielded many new insights (Wang, Gerstein, and Snyder 2009). mRNA-seq provides experimental evidence for genome annotation, enabling identification of differentially-spliced transcript isoforms, and previously-unrecognized coding and non-coding RNAs, as described in Chapter 2 (Mudge and Harrow 2016), while also providing a comprehensive, quantitative picture of transcript diversity. Whole transcriptome profiling of various cells and tissues, under diverse conditions, permits the identification of tissue / lineage / stage / strain-specific signatures (Uhlén et al. 2015), which can be used to construct metabolic and physiological models (de Oliveira Dal'Molin et al. 2016), examine complex samples such as pathogen-infected tissues (Westermann, Gorski, and Vogel 2012), *etc.*

Although comparing global gene expression patterns was not the primary goal of this study, we have generated and sequenced 24 strand-specific RNA libraries from several developmental stages/strains, as noted above, and examined these in conjunction with 45 additional libraries (Table 1). In aggregate, this provides the most comprehensive collection of data available on *T. gondii* transcript abundance across the parasite life cycle. Overall, transcriptomic results are highly consistent with biological and phenotypic variation across the complex parasite life cycle (Fig 1), including previously undescribed differences in gene expression during intracellular tachyzoite replication. Strong circumstantial evidence also suggests that lncRNAs may play an important role in regulating stage-specific expression during sexual differentiation and sporogony.

Methods

Data analysis and visualization

FPKM values for all genes (8,920 rows in Additional File 3) in the *T. gondii* genome (ToxoDB version 28) were filtered to restrict to those with FPKM value ≥ 20 (5,645 genes) for PCA and MDS analysis using the statistical excel add-in XLSTAT (Addinsoft; Figs 3.1, 3.2). Pearson (n) standardization was used for PCA, between-sample proximity matrices were calculated from the factor scores based on Euclidean distances. For MDS, we used the proximity matrix from PCA factor scores and selected the absolute model, Kruskal's stress, random initial configuration, 10 repetitions and stopped conditions if convergence was equal to 10^{-5} with > 500 iterations.

Results

Stage-specificity of *T. gondii* gene expression

RNA-seq datasets in this study were exploited to examine how global gene expression patterns compare across the parasite life cycle, from the acutely lytic tachyzoites responsible for disease, to latent bradyzoite tissue cysts in the brain, to sexual stage gametocytes from feline intestinal epithelium, to unsporulated oocysts in the environment (the fertilized zygote), and finally to the sporulated meiotic progeny (sporozoites) that are the immediate precursors to reinitiation of tachyzoite infection (Dubey, Lindsay, and Speer 1998). Additional File 3 presents FPKM values for the entire *T. gondii* genome for the 20 high quality strand-specific RNA-seq experiments highlighted in Table 1. Sample-specific correlations were performed using all genes that are expressed at FPKM >20, in any sample (63% of the genome).

Principal component analysis (PCA) clearly separates major life cycle stages, correlating well with parasite biology (Table 3). PC1 accounts for 27% of variation, and readily distinguishes oocysts (both sporulated & unsporulated) from gametocytes from asexual stages (both tachyzoites & bradyzoites). PC2 (23%) distinguishes gametocytes and unsporulated oocysts from tachyzoites/bradyzoites & sporozoites (see Fig 3.1 for a plot of PC1 *vs* PC2). PC3 (12%) distinguishes sporulated *vs* unsporulated oocysts; PC4 (11%) appears to distinguish early *vs* late stage tachyzoites. PC5 (8%) distinguishes mature bradyzoites from other stages. Overall, the top 10 principal components explain ~97% of observed variation between the 20 samples included in this study.

Sample	PC 1 *26.7%*	PC 2 *23.1%*	PC 3 *11.7%*	PC 4 *11.4%*	PC 5 *7.9%*	PC 6 *5.2%*	PC 7 *3.2%*	PC 8 *2.5%*	PC 9 *2.2%*	PC 10 *1.9%*
2Tz4_S	2.19	-7.25	-6.61	-30.09	-6.83	-7.90	9.04	-8.55	1.59	7.64
2Tz8_S	0.10	-2.53	-10.32	-40.55	-6.02	9.59	9.17	-11.19	-2.37	11.83
2Tz16_S	8.98	-6.85	-13.13	-23.12	-6.66	16.35	5.56	-6.32	-4.15	14.02
2Tz36_S	17.30	-24.59	-20.85	24.36	-10.83	15.77	-14.32	-0.59	-4.53	22.11
2Tz44_S	-6.02	-39.66	-6.37	28.90	-7.18	-31.17	-7.82	-16.98	5.77	10.07
3Tz2_S	-12.63	-16.70	0.38	-31.60	-6.33	-19.12	-0.25	3.57	-0.65	-13.73
3Tz4_S	-2.51	-9.89	-4.36	-29.76	-5.55	-14.40	6.65	7.10	-1.62	-9.72
3Tz8_S	-6.73	-2.59	-6.15	-39.26	-7.96	1.29	1.27	2.64	-4.23	-5.12
3Tz16_S	13.63	-11.03	-12.92	-11.98	-5.74	11.25	4.34	16.31	-4.77	-6.47
3Tz36_S	12.28	-30.15	-19.75	26.70	-10.04	5.90	-16.26	19.07	-7.39	-1.76
3Tz44_S	2.45	-32.92	-12.67	16.52	-6.96	-11.99	-14.15	14.79	-5.28	-8.21
TzHl_S	32.43	-17.53	-20.24	31.37	7.61	43.44	20.63	-10.34	9.21	-17.42
2Bz_S	-7.52	-37.88	-4.15	26.31	7.33	-23.26	9.93	-19.12	10.88	-10.62
3Bz_S	3.09	-15.04	14.70	-8.85	88.10	0.27	-6.81	2.95	-6.46	4.41
3dGmt_S	47.67	42.95	17.16	-13.70	-1.23	-2.71	-10.76	8.95	29.93	5.03
5dGmt_S	57.46	71.09	22.67	17.60	-7.87	-9.24	-0.86	-14.92	-30.61	-6.23
7dGmt_S	64.73	56.80	20.19	14.50	-1.92	-6.05	-0.06	5.77	11.79	0.72
Sz0_S	-100.83	95.82	-54.20	16.31	9.16	-1.18	-3.15	0.18	3.29	-0.85
Sz4_S	-71.11	-8.19	64.70	-5.59	-14.00	27.07	-26.84	-12.17	2.74	-7.84
Sz10_S	-54.96	-3.87	51.91	31.93	-7.08	-3.93	34.69	18.85	-3.15	12.15

Table 3. Analysis of stage-specific gene expression patterns in *T. gondii* by Principal Component Analysis.

Principal Component Analysis (PCA) for the 20 high quality strand-specific RNA-seq samples defined in Table 1, using all genes that are ever expressed at FPKM >20, in any sample. Contributions of each sample to the first 10 principal components, accounting for ~97% of observed variation.

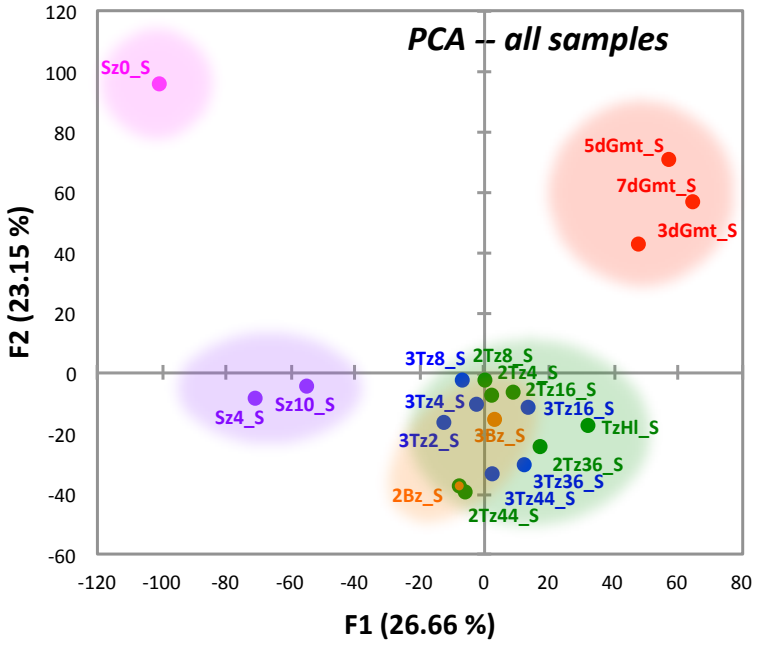


Figure 3.1. Spatial distributions of stage-specific gene expression patterns by PCA analysis. First two principal components for all samples, colored to highlight tachyzoites (*green*), bradyzoites (*orange*), gametocytes (*red*), unsporulated oocysts (*magenta*) & sporulated oocysts (*purple*).

See Additional File 3 for further details including eigenvalues, sample correlations, lists of specific genes making the greatest contribution to these principal components, eigenvectors for all genes, and the results of other analyses (K-means clustering, agglomerative hierarchical clustering, hierarchically-clustered expression heat maps, *etc*). Little is known about the specific genes or mechanisms involved in sexual differentiation in *Toxoplasma,* so it is difficult to compare these results with prior knowledge, but bradyzoite differentiation has been studied extensively (Dzierszinski et al. 2004; Soete, Camus, and Dubremetz 1993; Singh, Brewer, and Boothroyd 2002), and it is reassuring to find that major contributors to PC5, which clearly distinguishes mature bradyzoites from other stages, include the known bradyzoite differentiation markers BAG1, LDH2, ENO1, SRS9, SRS35A & SRS35B (Wasmuth et al. 2012; Dzierszinski et al. 2001).

Fig 3.2 uses multidimensional scaling (MDS; Hout, Papesh, and Goldinger 2013) to more accurately reflect between-sample distances in two dimensions (although the axes themselves are meaningless), particularly for various strains and stages of tachyzoites, which are not well-distinguished by PC1 & 2. Note that all 12 tachyzoite samples group together (green cloud), despite representing three strains, including the canonical type II & III lineages (*green & blue dots,* respectively; Howe and Sibley 1995; Grigg et al. 2001), and various early to late time points during intracellular replication. Additional analysis (not shown) demonstrates that type I strain tachyzoites co-cluster as well. *In vitro* bradyzoites (2Bz_S) cluster with tachyzoites (although in the direction of mature bradyzoites isolated from murine brain cysts; 3Bz_S, *orange*), as expected given the long, slow process of tachyzoite-to-bradyzoite differentiation (Dzierszinski et al. 2004; Soete, Camus, and Dubremetz 1993).

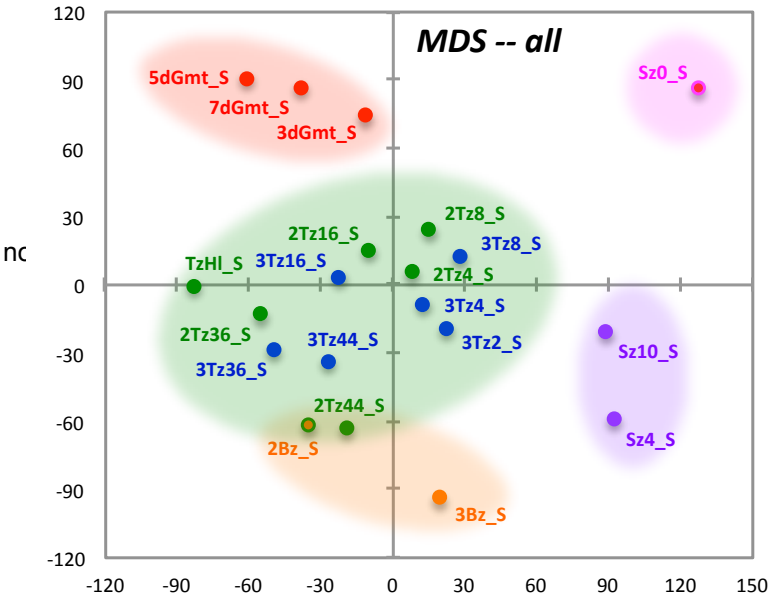


Figure 3.2. Spatial distributions of stage-specific gene expression patterns by MDS analysis.
MDS of the same samples; that for tachzyoites, samples group based on time post-infection (hr 2, 4, 8 …) rather than parasite strain (2Tz=ME49, 3Tz=VEG)

Gametocytes (*red*), unsporulated oocysts (*magenta*), and sporulated oocysts (*purple*) are also readily distinguished, justifying the binning of these samples, as shown

in Fig 3.3 (*green and blue dots* represent tachyzoites samples binned by strain; black dot represents pooled data for all *T. gondii* tachyzoite samples).

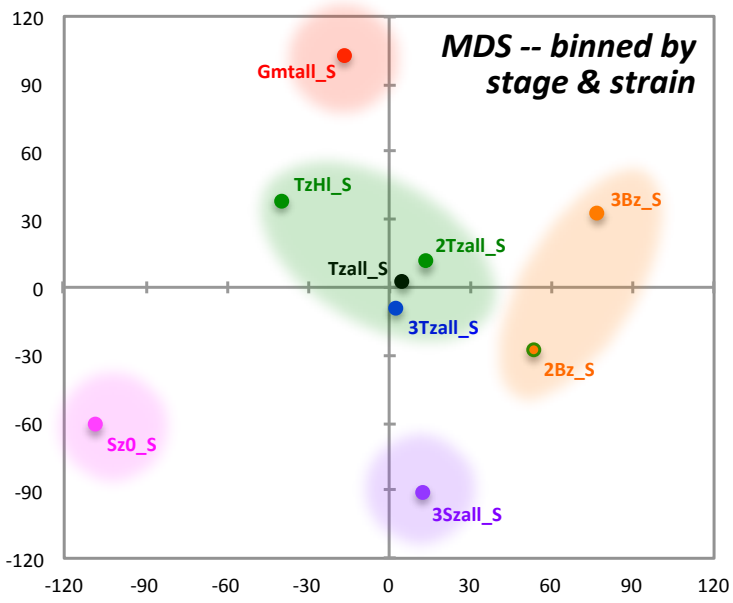


Figure 3.3. Spatial distributions of stage-specific gene expression patterns by MDS analysis binned by stage and strains. MDS for samples binned samples by stage and strain.

It is interesting to note that tachyzoite samples from similar time points post-infection are often more similar to each other than early *vs* late time points from the same strains in Fig 3.2, *i.e.* the 8 hr post-infection samples of ME49 *vs* VEG parasites (2Tz8 & 3Tz8) are relatively similar to each other, as are 36 & 44 hr time points. In contrast, ME49 strain 8 & 36 hr time points are quite distinct, as are VEG strain 8 & 36 hr time points. PCA analysis of samples binned based on time post-infection (Fig 3.4) shows that PC1 & 2 account for >76% of the observed variance.

Principal component 1 (PC1) does a good job of stratifying samples based on early *vs* late infection, and the top indicators are slightly enriched in genes associated with intracellular trafficking ($P<10^{-3}$); PC2 correlates with intracellular *vs* extracellular tachyzoites, and is enriched in genes bearing GO terms associated with redox balance and

chromatin assembly ($P<10^{-3}$). Functional distinctions over the course of intracellular tachyzoite infection have not been extensively studied, but analysis of genomic-scale expression profiles provides unprecedented resolution. These results also resonate with anecdotal observations of other differences, such as changes in splice site selection during the course of intracellular growth (Fig 2.9).

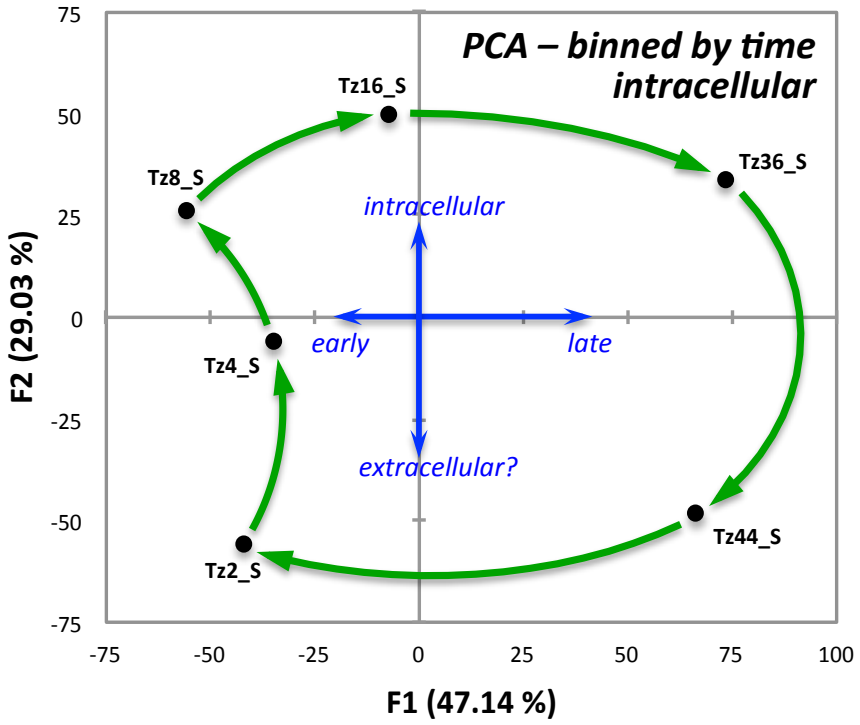


Figure 3.4. Spatial distributions of tachyzoite-specific gene expression patterns by PCA analysis binned by time post-infection.

PCA for tachyzoite samples binned based on time post-infection.

A role for opposite strand transcripts in regulating stage-specific expression?

As noted above, in addition to protein-coding genes, long non-coding transcripts (lncRNAs; Patil et al. 2013) are also observed *T. gondii* RNA-seq datasets, and circumstantial evidence suggests a possible role in transcriptional regulation. For example, the location of tachyzoite-specific lncRNAs in Figs 2.3 & 2.11A (*green boxes*) suggests that

they may act to suppress (or degrade) full-length mRNAs on the opposite strand. A quick search of the *T. gondii* database for genes with abundant support for sense-strand transcription in tachyzoites and antisense transcription in gametocytes yields dozens of genes, including several AP2-family transcription factors (Walker, Gissot, Huot, et al. 2013; Walker, Gissot, Croken, et al. 2013; Grigg et al. 2001; Behnke et al. 2014; Hehl et al. 2015). As shown in Fig 3.5A, there is a strong inverse correlation between expression of *Tg*ME49_282210 (AP2-VIIA8) mRNA in tachzyoites but not gametocytes, and opposite strand transcription in gametocytes but not tachzyoites. A similar inverse relationship was also observed for mRNA expression in bradyzoites and sporozoites, and lncRNA transcription in oocysts (not shown, but see Additional File 4). Neither stage-specific differences nor opposite strand transcripts were detected in the adjacent gene (*Tg*ME49_282210), which is also a putative AP2 transcription factor (AP2-VIIA9). Further inspection reveals a similar sense and antisense transcription patterns for many other genes in the parasite genome (Fig 3.5B).

To further explore the relationship between mRNAs and opposite strand lncRNA expression genome-wide, sense *vs* antisense transcript abundance was examined for various pairwise sample comparisons. Comparing tachyzoites with gametocyte samples (Fig 3.6) shows that the vast majority of genes are similarly expressed (central *black* cloud), consistent with observations in Table 2A. Some genes display stage-specific expression, with steady-state levels that differ by up to 10^5-fold. Positive controls for tachyzoite-specific, gametocyte-specific, and constitutively expressed genes are indicated in *red, purple* and *green,* respectively). The size of datapoints are scaled to reflect maximum mRNA abundance for the comparisons shown (FPKM values), *i.e.* large dots reflect abundant transcripts (ribosomal RNAs are indicated in gray).

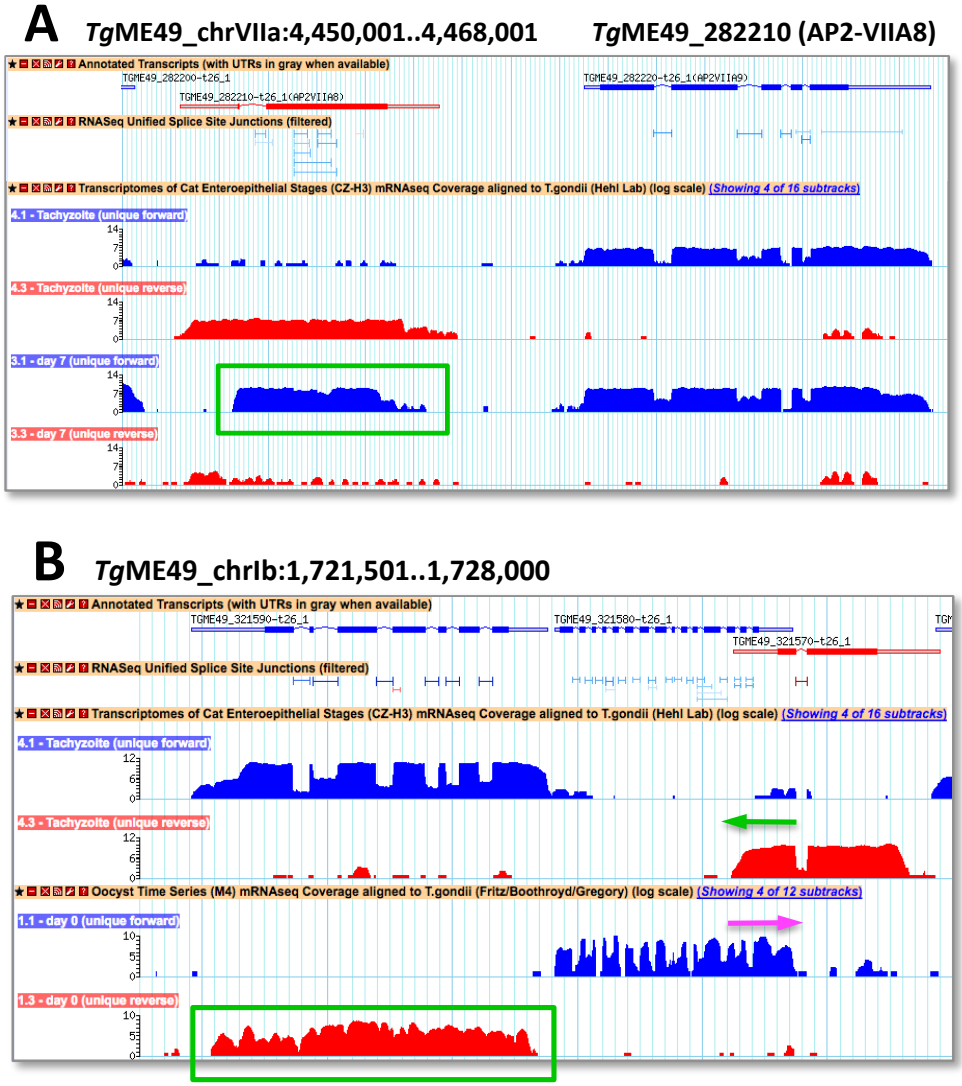


Figure 3.5. Selected examples of genes displaying an inverse stage-specific correlation between antisense RNA and mRNA.

Top: Gametocyte vs Tachyzoites. mRNA for *Tg*ME49_282210 (encoding AP2-VIIA8; *left*) is expressed on the reverse (*red*) strand in tachyzoites, while overlapping lncRNA is transcribed on the forward (*blue*) strand in gametocytes (see also reciprocal arrangement in Figs 2.3 & 2.11A). Note that the adjacent functionally-related gene (*Tg*ME49_282220 = AP2-VIIA9; *at right*) is constitutively expressed, with no antisense transcription. ***Bottom: Oocysts vs Tachyzoites.*** Tachyzoite-specific mRNA (*Tg*ME49_321590) and overlapping lncRNA on the reverse strand in gametocytes (*left*); the adjacent tachzyoite- and oocyst-specific mRNAs (*Tg*ME49_321570 & -321580, respectively) are convergently transcribed, with long overlapping 3'UTRs (*right*).

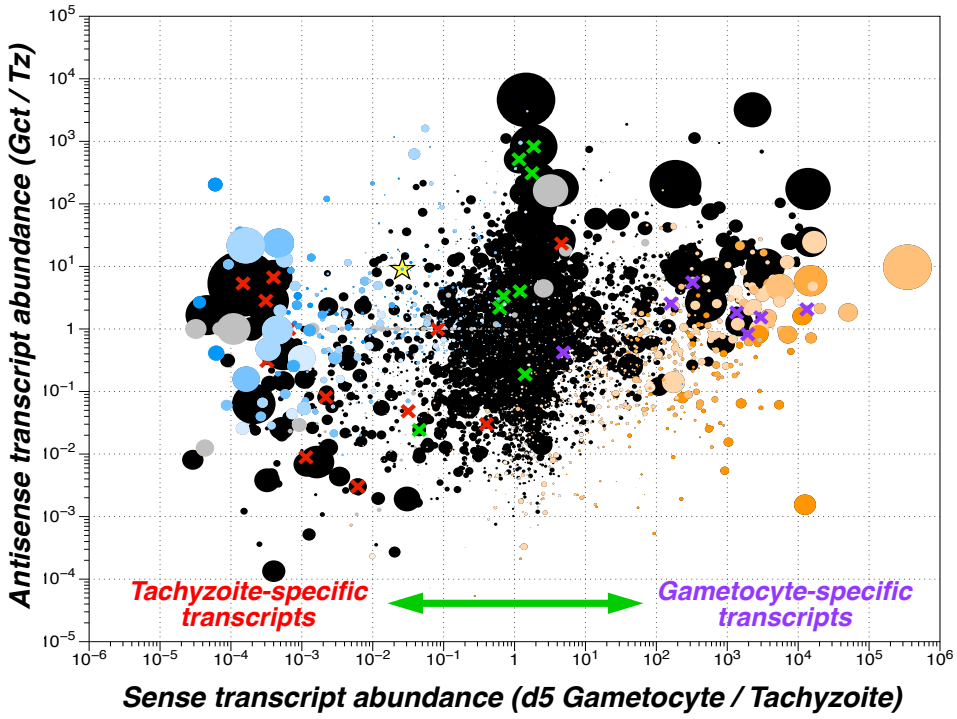


Figure 3.6. Antisense RNAs are inversely correlated with stage-specific mRNA transcript abundance during *T. gondii* differentiation.

The ratio of tachyzoite/gametocyte expression on the mRNA sense-strand (*horizontal*) *vs* anti-sense expression (*vertical*), for all annotated genes in the *T. gondii* genome. Points are scaled to reflect maximal mRNA abundance for the two samples represented in each graph; shading indicates excess antisense/sense strand expression in tachyzoites (*orange*) or gametocytes (*blue*); gray denotes ribosomal transcripts. For example, the yellow star at center/left corresponds to the transcription factor AP2-VIIA8 (*Tg*ME49_288210), which is abundantly transcribed on the sense strand in gametocytes, and overlaps a tachyzoite-specific lncRNA on the opposite strand (left part of Fig 3.5A); **X**'s indicate transcripts known to be constitutive (actin, alpha-tubulin1, histone H3.3, Asp1, calmodulin; *green*), or specific to tachyzoites (SAG1, GRA1, MIC1, ROP15, ENO2; *red*), gametocytes (SRS12A, SRS15A, SRS22B-1&2, SRS37B; *purple*)

Antisense transcription is generally less abundant than mRNA (*cf.* Fig 2.11), and less variable between samples, but differences of up to 10^3-fold are observed for some genes. In gametocytes, these is a slight correlation between sense and antisense expression, resulting in a positive overall slope, especially in the right half of Fig 3.6.

Shading reflects an overabundance of antisense transcript relative to the mRNA, *i.e.* tachyzoite-specific mRNA transcripts associated with gametocyte-specific antisense transcription (such as *Tg*ME49_282210; Fig 3.5A) are shaded blue.

Conversely, many gametocyte-specific genes (at right) are transcribed on the opposite strand in tachyzoites (*orange shading*). These differences are even more evident in comparisons of tachyzoites with parasite oocysts (Fig 3.7). More genes are differentially transcribed in tachyzoites *vs* unsporulated oocysts, and a slightly wider range of antisense transcript ratios is also observed (there was no discernable correlation between sense and antisense transcription within these tachyzoite or oocyst experiments) . More tachyzoite-specific genes display excess antisense transcription in oocysts, and the majority of unsporulated oocyst-specific transcripts are correlated with tachyzoite-specific antisense transcription. Three intriguing examples are highlighted in Fig 3.5B: *Tg*ME49_321590 (an uncharacterized 'hypothetical protein') is transcribed on the mRNA (sense, *blue*) strand in tachyzoites (as well as bradyzoites & gametocytes; not shown), and on the opposite (antisense, *red*) strand in unsporulated oocysts (and sporozoites; not shown). The adjacent genes *Tg*ME49_321580 (a putative membrane protein) and *Tg*ME49_321570 (FabZ) are also stage-specifically expressed, in either unsporulated oocysts only (*Tg*ME49_321580) or tachyzoites (*Tg*ME49_321570). Interestingly, these genes are convergently transcribed, with 3' UTRs that overlap extensively, providing the potential for antisense-mediated regulation based on mRNAs rather than lncRNAs.

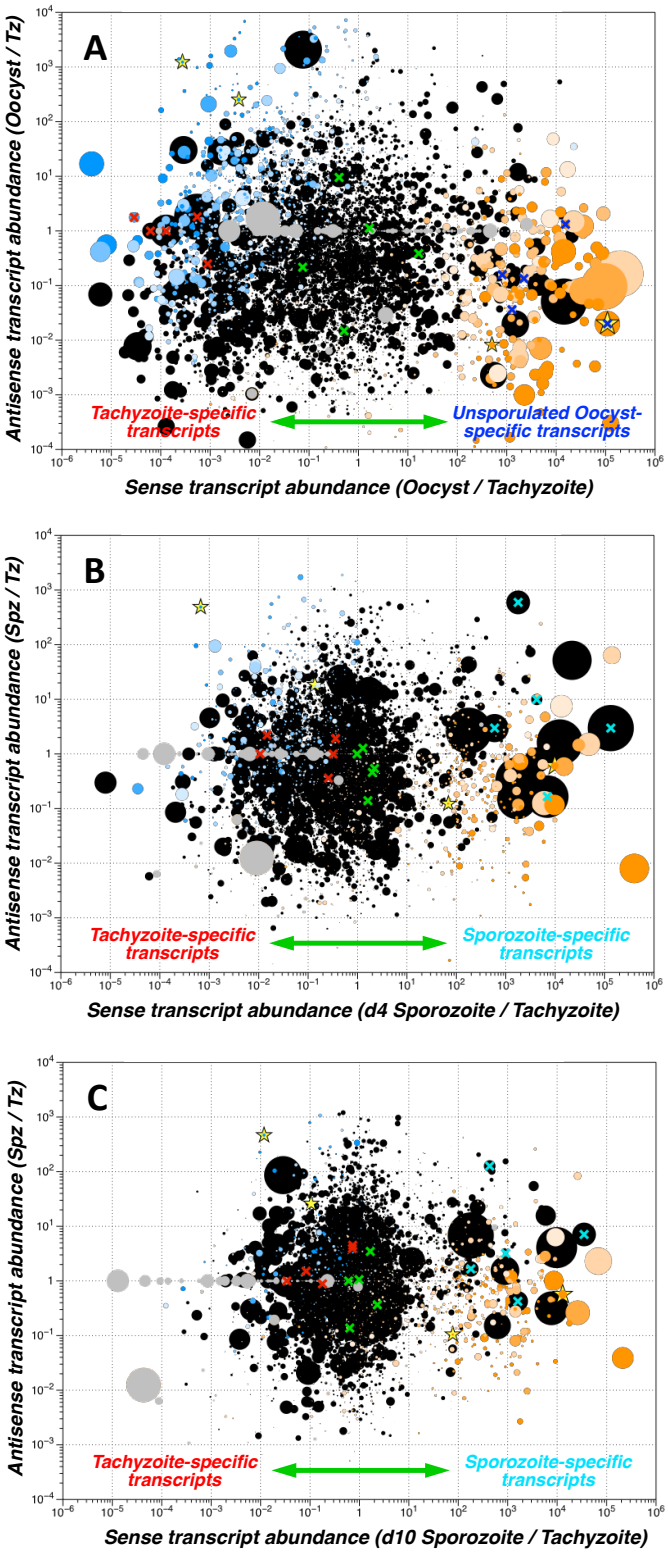


Figure 3.7. Antisense RNAs are inversely correlated with stage-specific mRNA abundance during *T. gondii* differentiation.

(**A**) tachyzoites *vs* unsporulated oocysts, (**B**) d4 sporulated oocysts, (**C**) d10 sporulated oocysts. Points are scaled to reflect maximal mRNA abundance for the two samples represented in each graph; shading indicates excess antisense/sense strand expression in tachyzoites (*orange*) or oocysts (*blue*); gray denotes ribosomal transcripts. For example, the yellow star at far right in panel A corresponds to Oocyst Wall Protein 2 (*Tg*ME49_209610), which is abundantly transcribed on the sense strand in oocysts, and overlaps a tachyzoite-specific lncRNA on the opposite strand; other yellow stars correspond to the three genes in Fig 3.5B. '**X**'s indicate transcripts known to be constitutive (actin, alpha-tubulin1, histone H3.3, Asp1, calmodulin; *green*), or specific to tachyzoites (SAG1, GRA1, MIC1, ROP15, ENO2; *red*), oocysts (OWP2, TBPIP, RAD54, *Tg*ME49_249520 & 263010; *blue*), or sporozoites (MIC13, SRS28, *Tg*ME49_209920, 259900 & 276880; *turquoise*).

Samples isolated after the induction of sporulation provide the opportunity to follow changes in sense and antisense transcript abundance during differentiation. Day 4 sporozoites (Fig 3.7B) are much more similar to tachyzoites, and day 10 sporozoites even more so (Fig 3.7C). *i.e.* the cloud of transcription ratios becomes progressively tighter. Importantly, antisense transcripts also change in parallel with this change in transcriptional profile, as oocysts differentiate into the sporozoites that will transform into tachyzoites once they infect a host cell: orange dots become less prominent, as oocyst genes (with tachzyoite-specific antisense transcripts) are downregulated, and the blue dots essentially vanish, as the antisense RNAs to tachyzoite-specific mRNAs that were abundant in oocysts disappear during the course of sporozoite differentiation.

Discussion

At least 69 RNA-seq samples are currently available for *T. gondii* parasites, many of which are described here for the first time (Table 1). Considered in their entirety, these studies sample strain diversity (Hassan et al. 2012), the intracellular tachyzoite replicative time course, and most major life cycle stages. Expression profiling has been used to investigate developmental signatures in many systems (Uhlén et al. 2015), and it is not surprising that major developmental stages can readily be distinguished (Table 3 & Figs 3.1 -3.4). Results obtained from these analyses are consistent with previous analyses of individual genes, and array-based transcriptional profiling (Fritz, Buchholz, et al. 2012; Wasmuth et al. 2012; Buchholz et al. 2011; Radke et al. 2005; Bahl et al. 2010). Given their substantial biological differences, it is gratifying to observe that differences between life cycle stages are much larger than differences between strains (Fig 3.2). The relatively low cost of RNA-seq allows analysis of multiple samples, including time

course studies, and the high degree of resolution provided by strand-specific RNA-seq identifies reproducible differences even within the 48 hr intracellular tachyzoite replicative cycle, which has not previously been reported (Fig 3.3, Fig S6). In this regard, it is interesting to observe consistent differences in transcript splicing (Fig 2.9B, Fig S6), although the biological significance (if any) of these isoforms is not known.

It is also clear that transcripts need not be restricted to mRNAs, as hundreds of lncRNAs are evident in the *T. gondii* genome, often overlapping stage-specific mRNAs, and reciprocally expressed (Figs 2.3, 2.11A & 3.5). Some lncRNAs span the opposite strand gene (Fig 3.5), while others cover only a small portion (Figs 2.3 & 2.11A). In other cases, stage-specific expression of convergently-transcribed genes yields long 3'UTRs providing antisense coverage (Fig 3.5B, *right*). Inversely correlated sense and antisense transcripts are most prominent in comparisons of tachyzoites with oocysts (Fig 3.7), followed by tachyzoites *vs* gametocytes (Fig 3.6). There is little evidence for antisense association with the developmental transition from tachyzoites to bradyzoites (not shown, but see Additional File 4). The inverse correlation between sense and antisense RNA abundance suggests a causal relationship, and it is tempting to rationalize these observations by imagining that the tight regulation of transition to gametocytes (in the cat gut), and oocysts (whose metabolism is very different, and must survive for months in the harsh external environment), could explain the use of non-transcriptional mechanisms for silencing. Evaluating such hypotheses will require experimental testing, ideally using an *in vitro* model of the coccidian cycle in the definitive host.

The nature of lncRNAs is not entirely clear in *T. gondii* and other apicomplexan parasites: PCR amplification of putative lncRNAs has not always been reliable, suggest-

ing the possibility of multiple shorter transcripts rather than one long transcript; upstream sequences are not always effective as promoters in reporter assays; lncRNAs are not generally associated with active or repressive chromatin marks; and it is not even certain that *T. gondii* lncRNAs are all polyadenylated Pol II transcripts. *T. gondii* does possess Argonaut and Dicer proteins, but it is not clear that these mediate silencing (Meissner et al. 2007; Crater et al. 2016). It would be very helpful to have more effective methods for lncRNA prediction, but such algorithms have proved difficult to develop, and are not broadly applicable across diverse species (Fiscon, Paci, and Iannello 2015; Liao et al. 2011). By definition, coding potential is not relevant for assessing non-coding RNAs, and many lncRNAs display evidence of multiple introns that are very inefficiently excised (*cf.* Fig 3.5).

CHAPTER 4: SUMMARY, GENERAL DISCUSSION AND FUTURE DIRECTIONS

Structural annotation of gene models has been a challenge since the early days of genome sequencing, and has engaged both molecular biologists using experimental methods (RT-PCR, EST sequencing) and computational biologists responsible for *ab initio* and *de novo* methods. Deep sequencing of RNA molecules has revolutionized gene prediction, as it is now possible to obtain hundreds of millions of short RNA (or polyA+ mRNA) reads from individual samples. RNA-seq reveals an immense number of splice junctions, however, making it difficult to distinguish functional mRNAs from splicing machinery errors and experimental artifacts. Distinguishing alternatively spliced isoforms that that are likely to be functional and should be annotated, from those that should not, is an important unsolved problem.

In order to study the prevalence of alternative splicing, we analyzed transcript expression from all available RNA sequencing experiments for the protozoan parasite *Toxoplasma gondii,* which displays typical eukaryotic splicing. This species' genome is substantially smaller than that of humans and most common model organisms: ~65Mb, encoding 8322 protein coding genes, with 40,103 annotated introns. Several alternatively-spliced genes have previously been characterized in *T. gondii,* including examples of intron retention, exon-skip polymorphisms, and splice acceptor and donor variants … providing positive controls for this analysis (Fig. 2.1). There is also good reason to believe that algorithms used to identify alternative splice products in *T. gondii* will be applicable to other systems, as human nuclear extracts can properly splice *T. gondii* primary mRNAs, and vice versa.

Gene annotation

Taken together, this work outlines effective strategies for exploiting RNA-seq datasets to improve the annotation of draft eukaryotic genomes. New RNA sequencing technologies provide high depth, high quality transcriptomic evidence from entire genome(s), which has proven to be extremely valuable for improving annotation, and revealing unappreciated transcripts (Djebali et al. 2012; Harrow et al. 2012; Mudge and Harrow 2016).

The aggregate pool of RNA-seq data for all samples initially considered in this study (Table 1) provides considerable information: the observed total of >2.7 million reproducibly-observed introns expands the number of *T. gondii* transcripts by many fold. Most of these junctions are exceedingly rare, however, as they were observed fewer than six times in the total of 69 possible samples. Furthermore, many of these rare junctions usually lie in non-coding regions such as UTRs (Fig 2.1) and lncRNAs and this is not surprising as higher sequence variation is known to be present in less conserved non-coding sequences than in coding sequences (Mu et al. 2011; Castle 2011).

The current *T. gondii* genome annotation was used as a guide to define parameters that minimize false discovery of transcripts. Exploiting the relationship between the abundance of reads that span introns or ISRPM and the reads that map to annotated genes or FPKM (to which the introns map to), provided good separation between True Positive (*black* datapoints in upper half of Fig. 2.6) and True Negative introns (black numbers in the right-hand column of Table 2A), even when applied to other life cycle stages (Table 2A and Fig 2.7A) and other species (*Neospora, Plasmodium;* same parameters were applied to genomes in EuPathDB.org). Introns with low ISRPM/FPKM are

inefficiently spliced and can be excluded from further analysis even if they are frequently observed in heavily transcribed genes. An even better strategy is to directly compare intron abundance to the number of reads in the flanking exons of each intron, instead of using the expression of the entire gene to which the intron maps to. Several other methods use ψ or Percent Spliced In to study the prevalence of alternative splicing across transcriptomes (Venables et al. 2008; Katz et al. 2010; Shen et al. 2014; Trapnell et al. 2013; Vaquero-Garcia et al. 2016). These alternative methods are aware of the change in coverage that can occur in a gene, if the gene is misannotated for example, and that is usually masked out by the fact that the total number of reads that map to a gene are normalized by the length of the gene, even if the coverage is not uniform in different exons of the same gene. Despite these differences, the method described here serves as an adequate heuristic when existing genome annotation is close to accurate. This method is also relatively straightforward to implement, as it is relatively straightforward to automatically extract from RNA-seq mapping pipelines and genome databases. This method could be easily applied to other organisms for which similar databases are available.

We found that best proportion of TP to TN introns occurred when we considered the maximal ISRPM from the 20 high quality samples selected in this study (Table 1) and compared it to the FPKM of the same sample where the maximal ISRPM was expressed (Table 2A and Fig 2.7B). This analysis showed that most False Positive introns individually analyzed in different samples turned out not be real FP introns: many were stage-specifically expressed (Fig 2.7B) and some were real alternatively-spliced isoforms (Fig 2.8). Likewise numerous False Negative introns were not completely false as they were

alternatively-spliced, but were less abundant than the annotated TP alternative (Table 2B).

In summary, we show that >93% of ~40K annotated introns associated with ~9K genes in the reference genome are strongly supported by RNA-seq data from at least one sample, leaving ~5% that should be deleted, and 2% that should be supplanted by alternative introns. In addition, we describe ~20% well-supported new introns, including 4% that are alternatively-spliced (many within UTRs), ~8% that modify existing gene models, and ~8% associated with novel transcripts (approximately equally divided between predicted mRNAs and lncRNAs). Abundant introns associated with novel transcripts represent either genes that are specifically expressed in previously under-studied stages (*cf.* Figure 2.5) or lncRNAs whose exact function is not known, but thought likely to be involved in transcriptional regulation (*cf*. Figs. 3.5-7). One of the difficulties in the current annotation is the definition of UTR boundaries, as the algorithms that were used for calling UTRs applied RNA-seq data in a greedy manner. Instead of using the most abundant transcript to define UTR boundaries, the algorithm incorporated information from the longest possible transcript. This means that a portion of annotated gene models have potentially long UTRs, sometimes affecting correct call of transcriptional and/or translational start sites (*cf*. Figs 2.7-9). Usage of the most abundant transcript from RNA-seq data instead of the longest, should correct this problem in principle. However, as shown in Figure 2.8, this solution is difficult to implement as a general rule in all annotated genes. For example, in the case of Isocitrate Dehydrogenase (Fig 2.8) it would result in exclusion of the first annotated intron, which is clearly excised more frequently in a stage-specific manner in tachyzoite parasites. This intron is less often excised in the cat enteroepithelial stages because there is a stronger promoter within

this intron that would preclude excision most of the time. It is difficult to decide which transcripts should be annotated and which ones should not.

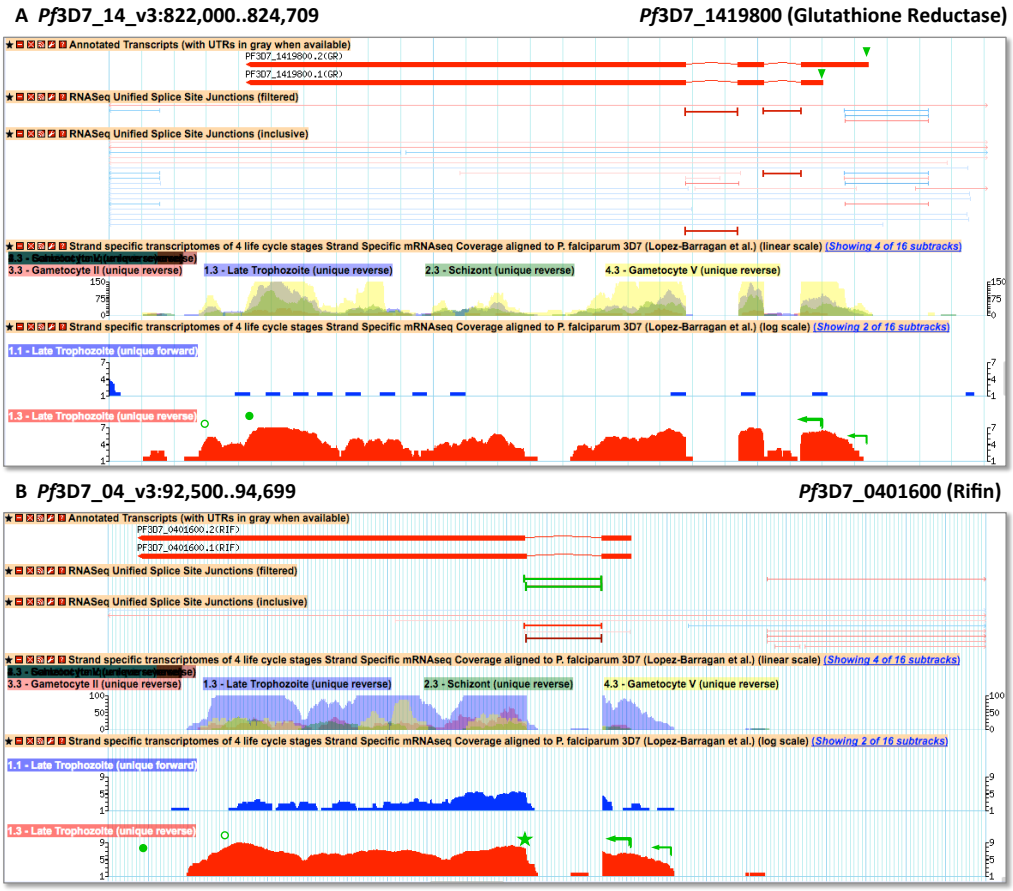


Figure 4.1. Application of ISRPM/FPKM parameters to *Plasmodium* RNA-seq datasets discriminates significant introns from low abundance variants (from PlasmoDB.org). Genome annotation and selected RNA-seq tracks from gametocytes and late trophozoites. Note that blue and red tracks are scaled logarithmically, while tracks shown in semi-transparent overlay are linear (note different scales, selected to highlight stage specific expression). *Arrows* and *circles* indicate transcript termini. Note the two translational starts in A (arrowheads), which have been confirmed to have different localizations within the strain 3D7 of *P. falciparum* (Kehr et al. 2010). Two tracks display candidate introns (brackets): the 'filtered' track is restricted to display only those introns that pass the abundance and efficiency criteria described in this dissertation. The *star* in panel B indicates an acceptor splice variant that deletes a single aminoacid from the Rifin transcript.

Most biologically-significant errors in the current reference annotation for *T. gondii* relate to inaccurate annotation of transcript initiation, often impacting the predicted translational start site (Figs. 2.7 & 8), and hence subcellular localization predictions (which are particularly important for pathogenic species that secrete factors modifying the host response). These kinds of errors can be addressed by several mechanisms, such as implementing known features of transcriptional initiation context (Matrajt et al. 2004), incorporating evidence from conserved sequences in orthologous genes and proteins from other related species as well as evidence from proteomic experiments. Other solutions could also be to consider data from promoter binding experiments (Niu, Tabari, and Su 2014) as wells as from full length RNA-seq (Sharon et al. 2013) and other sequencing technologies that sequence fragments of up to 150kb (Jain et al. 2016).

Longer reads will also benefit research and analysis of alternative splice variants, as complete or near complete sequenced RNAs would have less sources of error generally produced during the assembly of short reads from current Illumina NGS into transcript(s). In order to estimate the significance of alternative splice variants, phylogenetic comparisons aiming to ask if putative variants are conserved in several strains of *T. gondii* or in other closely related species, such as *Neospora sp*. and *Plasmodium sp*. should be considered. Examination by multiple sequence alignments of how different regions within potential variants change in sequence in their ortholog counterparts should provide more evidence towards the functional significance, if any, as well as the conservation of these variants among several related species. Characterization of changes in abundance of alternative transcripts across different spliceosomal mutants

should provide an extra layer to evidence about the potential biological significance of splice variants.

The existing genome annotation could also benefit from identifying new unannotated transcripts such as lncRNAs. Our current ability to predict plausible lncRNAs from sequencing data is poorly developed, as definition of lncRNA transcripts has proven difficult because they do not contain the same sequence features as protein coding transcripts. Developing new gene finders that use machine learning techniques to learn sequence features from known *T. gondii* lncRNAs would be the best approach to characterize lncRNAs, as methods developed for other species have proven not applicable in *T. gondii* since they are specific for the system for which they were developed (Liao et al. 2011; Fiscon, Paci, and Iannello 2015).

Transcriptional regulation

In order to study possible mechanisms of transcriptional regulation, we also carried out transcriptional profiling by RNA-seq, including other available transcriptional data to encompass sequencing information from all stages of the complex parasite life cycle (Table 1). Steady-state transcript levels highlight important differences, even within the intracellular replication cycle, and circumstantial evidence suggests an important role of antisense transcription (including lncRNAs) in regulating stage-specific expression. Stage-specific expression was previously known is readily distinguished in all major life cycle stages by PCA and MDS analysis (*cf.* Figs. 3.1-3). As expected, between-stage differences were greater than between-strain differences (Fig 3.2), and to our knowledge, this report provides the first observation of profiling differences within the tachyzoite replicative cycle (Fig 3.4), including differences between intracellular and

extracellular parasites, and early and late stage intracellular tachyzoites. It will be interesting to determine the functional significance of these differentially-regulated transcripts within the tachyzoite lytic stage.

Analysis of strand-specific RNA-seq data has identified a previously unappreciated diversity of (unannotated) long non-coding RNAs (lncRNAs). We looked for differential expression of sense vs antisense transcripts in tachyzoites (the acutely lytic stage causing disease, which is readily cultivated) vs either bradyzoites (the latent cyst stage, which is less readily grown in culture), gametocytes (tachyzoites or bradyzoites undergoing sexual differentiation in the epithelium of the cat's gut) or sporozoites (the result of sexual crosses, which is not readily obtained in the lab) to examine elucidate the relationship between sense and antisense transcripts. The function of these transcripts is not yet clear, but many overlap the 3' UTR of coding genes, and display stage-specific mutually-exclusive coding vs non-coding transcript expression (*cf.* Figs. 3.5, 3.6 & 3.7), suggesting a possible role in the regulation of stage-specific expression, specially during the sexual stages.

In order to determine if antisense transcripts play a significant role in transcriptional regulation during parasite biology, several strategies and experiments could be carried in future research. One would be to map antisense lncRNAs promoters in *T. gondii* tachyzoites using for example a luciferase transient transfection assay. Other experimental option to elucidate the function of antisense transcriptional regulation would be to test the hypothesis that antisense lncRNAs regulate expression of sense overlapping mRNA transcripts, by manipulating their expression in transgenic parasites. Three experimental approaches could be used to test this hypothesis. One would be using the Ku80 knock-

out system for targeted homologous recombination in *T. gondii* to delete the 5' upstream region (and ideally the promoter region) for tachyzoite-specific lncRNAs. Antisense and sense transcript expression will then be assessed by strand-specific qPCR. A second experimental strategy would be to overexpress antisense lncRNAs in tachyzoites from a tet-inducible promoter, following antisense and sense transcript expression/repression by strand-specific qPCR. Finally, a third experimental approach could exploit CRISPR mutagenesis to specifically change the sequence of strand-specific antisense transcripts and or their promoters in tachyzoite parasites and then monitor antisense and sense transcript expression/repression by strand-specific qPCR. These strategies, however, would generate fewer insights into a possible biological role of antisense transcripts because the most striking regulation appears to be occurring in the sexual stages of the parasite, especially sporozoites, where manipulation has proven to be difficult.

Other interesting projects could be continued from the data generated for this thesis. Sequencing experiments also yielded data on host cell responses to parasite infection as well as on small non-codingRNAs (sncRNAs) from the parasite and its host cell. Analysis of host cell responses to parasite infection should permit identification of co-regulatory patterns correlating host and pathogen gene expression, such as relation-ships between metabolic pathways and receptor-ligand interactions, among others. In related work, we also provided Jonathan Wastling's group at Liverpool with parallel samples for quantitative proteomic analysis, and it will be interesting to compare parasite and host transcriptomics vs proteomics.

Mapping of sncRNAs reads to the parasite genome showed that they are most abundantly found on repeats and introns of genes (mRNA). Parasite expressed

sncRNAs are of several different sizes, varying from the 21-23nts size that are usually observed in humans. Different classes of sncRNAs, regulated in a tissue and time specific manner in plants and animals, are well known to play important roles in mechanisms such as defense, development and chromatin regulation (Kim and Sung 2012; O'Connell, Rao, and Baltimore 2012). However the apparent lack of an abundant unified size class (>50%) in *T. gondii* sncRNAs, indicates that the origin and possible function of sncRNAs is different from what is known in other organisms. The *T. gondii* genome codes for a Dicer and an Argonaute protein. The *Tg*Dicer protein, in charge of cleaving long dsRNA and generating siRNA or miRNA precursors, is composed of a RNA helicase domain and two RNase II catalytic domains. Nevertheless it lacks the conserved dsRNA binding domains, DSRM and PAZ, features also seen in the Dicer of the single cell Algae, *Chlamydomonas reindhartii*. The *Tg*Argonaute protein, displays conserved PAZ and PIWI domains, along with RGG amino terminal repeats (3X) which have the potential to be methylated, a characteristic shared with metazoan and plant Argonautes (Braun et al. 2010). *T. gondii* also possesses one representative of a RNA dependent RNA polymerase (RdRP) which is implicated in other organisms of amplifying sncRNAs from endogenous transcripts (Nishikura 2001). Modified tachyzoite parasites have been generated to contain a KO version of Argonaute, but standard characterization experiments don't seem to show any defect, so it is not clear what function *Tg*Argonaute is playing. To assess the possible function of sncRNA in *T. gondii*, it would be interesting to generate KO sexual parasites lacking each of the representatives possibly involved in small ncRNA biogenesis and test if they have any effect in the biogenesis and/or stability of small ncRNAs.

Human small ncRNAs were on average 23nt long, and mapped to several locations throughout the human genome, being most abundant in repetitive regions. The specific size of 23nt agrees with known animal miRs and siRNA sizes (McManus and Sharp 2002). Out of 317 human miRs analyzed, there were 164 (51.7%) that were suppressed two fold or more during infection with both *Tg*RH and *Tg*ME49. Only 18 (5.7%) of human miRs were upregulated two fold or more during infection with these two strains, among which we found and confirmed what has been previously reported: upregulation of members of the family miR-17 (Zeiner et al. 2010). Additionally we have also encountered that human miRs -146a and -155 are also among the upregulated microRNAs. These have been reported to mediate a negative feedback loop regulation of the human transcription factor NF-κB (Taganov et al. 2006), which during *T. gondii* infection is known to be activated in varying degrees by different strains (Rosowski et al. 2011). Examining *Toxoplasma*-infected human cells, we have now found that the more notable impact of parasite infection is a profound suppression of most human miR expression. Human miRs appear to be suppressed by infection with *T. gondii* strains RH and ME49, but not VEG and by *N. caninum*. It will be interesting to following up on these preliminary results, as they promise to reveal mechanisms of host-parasite interactions.

In summary, exploiting multiple RNA-seq datasets, from diverse sources, and integrating this information with other functional genomics data, permits substantial improvement in the annotation of the *Toxoplasma gondii* genome and transcriptome. The methods described here should be readily incorporated into computational pipelines, and are likely to prove useful for other species as well.

APPENDIX MATERIALS

The following supplementary materials are available in the supplemented CD:

Additional File 1: Microsoft Excel .xlsx file containing the following spreadsheets, each represented as a separate tab: Data, Stats, Annotated noEvidence (FP), Unannotated AltSplicing (FN) and Unannotated NewGenes (FN). 'Data' lists all annotated introns and unannotated intron-spanning reads observed at plausible abundance (59,755 rows), along with information on chromosomal position & strand, gene mapping & location within the gene, annotation status, ISRPM values, FPKM (when mapped to genes), and ISRPM/FPKM values, for all 20 high quality strand-specific RNA-seq datasets (see Table 1). Additional columns indicate the maximum ISRPM observed for each intron, the maximum ISRPM for any alternative overlapping intron, and corresponding FPKM & ISRPM/FPKM values. Appropriate columns from this table were used to generate Figs 2.4, 2.5, 2.6, 2.12 & 2.13. The 'Stats' tab provides counts and several statistics relating to the 'Data' tab, including data presented in Table 2. Other tabs present filtered views of 'Data' highlighting specific corrections recommended for the reference *T. gondii* genome annotation.

Additional File 2: PDF file containing supplementary figures S1-S5. Legends are included at the beginning of this file.

Additional File 3: Excel .xlsx file containing supplementary spreadsheets, presenting FPKM values for all 8920 annotated genes, from all 20 high-quality strand-specific

studies analyzed in this manuscript (Table 1), in addition to various averaged datasets, *e.g.* all strain ME49 parasites, all 4 hr post-infection samples, *etc*). Additional tabs present analysis of these data using various transcript abundance cut-offs, as noted, by PCA & MDS (see Figs 3.1, 3.2, 3.3 & 3.4 and Table 3), as well as HC (heatmap), AGC, *etc* (see Methods).

Additional File 4: Excel .xlsx file containing supplementary spreadsheets, presenting sense and antisense FPKM values for all 8920 annotated genes, from 20 high-quality strand-specific studies analyzed in this thesis (Table 1), plus the ME49 hr 2 sample, in addition to various averaged datasets, *e.g.* all gametocyte samples, and pairwise comparisons between samples *e.g.* the ME49 tachyzoite *vs* d7 CZ-H3 gametocyte comparison (presented in Fig 3.6). The second tab contains transformations of minimum and maximum FPKM values as indicated in the tab name.

BIBLIOGRAPHY

Ajioka, James W, John C Boothroyd, Brian P Brunk, Adrian Hehl, Ledeana Hillier, Ian D Manger, Marco Marra, et al. 1998. “Gene Discovery by EST Sequencing in Toxoplasma Gondii Reveals Sequences Restricted to the Apicomplexa.” *Genome Research* 8 (1): 18–28.

Andenmatten, Nicole, Saskia Egarter, Allison J Jackson, Nicolas Jullien, Jean-paul Herman, and Markus Meissner. 2013. “Conditional Genome Engineering in Toxoplasma Gondii Uncovers Alternative Invasion Mechanisms.” *Nature Methods* 10 (2): 125–27. doi:10.1038/Nmeth.2301.

Bahl, Amit, Paul H Davis, Michael Behnke, Florence Dzierszinski, Manjunatha Jagalur, Feng Chen, Dhanasekaran Shanmugam, Michael W White, David Kulp, and David S Roos. 2010. “A Novel Multifunctional Oligonucleotide Microarray for Toxoplasma Gondii.” *BMC Genomics* 11: 603. doi:10.1186/1471-2164-11-603.

Balaji, S, M Madan Babu, Lakshminarayan M Iyer, and L Aravind. 2005. “Discovery of the Principal Specific Transcription Factors of Apicomplexa and Their Implication for the Evolution of the AP2-Integrase DNA Binding Domains.” *Nucleic Acids Research* 33 (13): 3994–4006. doi:10.1093/nar/gki709.

Basso, Walter, Sonja Hartnack, Lais Pardini, Pavlo Maksimov, Bretislav Koudela, Maria C Venturini, Gereon Schares, Xaver Sidler, Fraser I Lewis, and Peter Deplazes. 2013. “Assessment of Diagnostic Accuracy of a Commercial ELISA for the Detection of Toxoplasma Gondii Infection in Pigs Compared with IFAT , TgSAG1-ELISA and Western Blot , Using a Bayesian Latent Class Approach.” *International Journal for Parasitology* 43 (7): 565–70. doi:10.1016/j.ijpara.2013.02.003.

Behnke, Michael S, Tiange P Zhang, Jitender P Dubey, and L David Sibley. 2014. “Toxoplasma Gondii Merozoite Gene Expression Analysis with Comparison to the Life Cycle Discloses a Unique Expression State during Enteric Development.” *BMC Genomics* 15: 350. doi:10.1186/1471-2164-15-350.

Belperron, Alexia A, Barbara A Fox, Toshihiro Horii, and David J Bzik. 2001. “Toxoplasma Gondii : Genetic Selection of Tethered Dihydrofolate Reductase – Thymidylate Synthase Fusion Proteins.” *Experimental Parasitology* 98 (3): 167–70. doi:10.1006/expr.2001.4631.

Bernal, Axel, Koby Crammer, Artemis Hatzigeorgiou, and Fernando Pereira. 2007. “Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction.” *PLoS Computational Biology* 3 (3): e54. doi:10.1371/journal.pcbi.0030054.

Bernal, Axel, Koby Crammer, and Fernando Pereira. 2012. “Automated Gene-Model Curation Using Global Discriminative Learning.” *Bioinformatics* 28 (12): 1571–78. doi:10.1093/bioinformatics/bts176.

Bernal, Axel E, and Fernando Pereira. 2012. “LINEAR STRUCTURE MODELS FOR EUKARYOTIC GENE PREDICTION in.” University of Pennsylvania.

Besteiro, Sebastien, Adeline Michelin, Joel Poncet, Jean-François Dubremetz, and Maryse Lebrun. 2009. “Export of a Toxoplasma Gondii Rhoptry Neck Protein Complex at the Host Cell Membrane to Form the Moving Junction during Invasion.” *PLoS Pathogens* 5 (2): e1000309. doi:10.1371/journal.ppat.1000309.

Blader, Ira J, Bradley I Coleman, Chun-Ti Chen, and Marc-Jan Gubbels. 2015. “Lytic Cycle of Toxoplasma Gondii: 15 Years Later.” *Annual Review of Microbiology* 69: 463–85. doi:10.1146/annurev-micro-091014-104100.

Bowman, Elizabeth A, and William G Kelly. 2014. “RNA Polymerase II Transcription Elongation and Pol II CTD Ser2 Phosphorylation: A Tail of Two Kinases.” *Nucleus* 5 (3): 224–36. doi:10.4161/nucl.29347.

Bradley, Peter J, Nancy Li, and John C Boothroyd. 2004. “A GFP-Based Motif-Trap Reveals a Novel Mechanism of Targeting for the Toxoplasma ROP4 Protein.” *Molecular and Biochemical Parasitology* 137 (1): 111–20. doi:10.1016/j.molbiopara.2004.05.003.

Braun, Laurence, Dominique Cannella, Philippe Ortet, Mohamed Barakat, Céline F Sautel, Sylvie Kieffer, Jérôme Garin, Olivier Bastien, Olivier Voinnet, and Mohamed-Ali Hakimi. 2010. “A Complex Small RNA Repertoire Is Generated by a Plant/Fungal-Like Machinery and Effected by a Metazoan-Like Argonaute in the Single-Cell Human Parasite Toxoplasma Gondii.” *PLoS Pathogens* 6 (5): e1000920.

Brooks, Carrie F, Maria E Francia, Mathieu Gissot, Matthew M Croken, Kami Kim, and Boris Striepen. 2011. “Toxoplasma Gondii Sequesters Centromeres to a Speci Fi c Nuclear Region throughout the Cell Cycle.” *Proceedings of the National Academy of Sciences* 108 (9): 3767–72. doi:10.1073/pnas.1006741108.

Buchholz, Kerry R, Heather M Fritz, Xiucui Chen, Blythe Durbin-johnson, David M Rocke, David J Ferguson, Patricia A Conrad, and John C Boothroyd. 2011. “Identification of Tissue Cyst Wall Components by Transcriptome Analysis of In Vivo and In Vitro Toxoplasma Gondii Bradyzoites.” *Eukaryotic Cell* 10 (12): 1637–47. doi:10.1128/EC.05182-11.

Burge, Christopher B, and Samuel Karlin. 1998. “Finding the Genes in Genomic DNA.” *Current Opinion in Structural Biology* 8 (3): 346–54.

Castle, John C. 2011. “SNPs Occur in Regions with Less Genomic Sequence Conservation.” *PLoS ONE* 6 (6): e20660. doi:10.1371/journal.pone.0020660.

Chaudhary, Kshitiz, Robert G K Donald, Manami Nishi, Darrick Carter, Buddy Ullman, and David S. Roos. 2005. “Differential Localization of Alternatively Spliced Hypoxanthine-Xanthine-Guanine Phosphoribosyltransferase Isoforms in Toxoplasma Gondii.” *Journal of Biological Chemistry* 280 (23): 22053–59. doi:10.1074/jbc.M503178200.

Chu, Ci, Qiangfeng Cliff Zhang, Simão Teixeira Da Rocha, Ryan A. Flynn, Maheetha Bharadwaj, J. Mauro Calabrese, Terry Magnuson, Edith Heard, and Howard Y. Chang. 2015. “Systematic Discovery of Xist RNA Binding Proteins.” *Cell* 161 (2): 404–16. doi:10.1016/j.cell.2015.03.025.

Crater, Anna K, Scott Roscoe, Mikayla Roberts, and Sirinart Ananvoranich. 2016. “Antisense Technologies in the Studying of Toxoplasma Gondii.” *Journal of Microbiological Methods*, pii: S0167-7012(15)30136-6. doi:10.1016/j.mimet.2015.12.013.

Croken, Matthew M, Sheila C Nardelli, and Kami Kim. 2013. “Chromatin Modifications, Epigenetics and How Protozoan Parasites Regulate Their Lives.” *Trends in Parasitology* 28 (5): 202–13. doi:10.1016/j.pt.2012.02.009.

Djebali, Sarah, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, et al. 2012. “Landscape of Transcription in Human Cells.” *Nature* 489 (7414): 101–8. doi:10.1038/nature11233.

Donald, Robert G K, Darrick Carter, Buddy Ullman, and David S Roos. 1996. “Insertional Tagging, Cloning, and Expression of the Toxoplasma Gondii Hypoxanthine-Xanthine-Guanine Phosphoribosyltransferase Gene. Use as a Selectable Marker for Stable Transformation.” *The Journal of Biological Chemistry* 271 (24): 14010–19.

Donald, Robert G K, and David S Roos. 1993. “Homologous Recombination and Gene

Replacement at the Dihydrofolate Reductase-Thymidylate Synthase Locus in Toxoplasma Gondii." *Molecular and Biochemical Parasitology* 63 (2): 243–53.

Dubey, Jitender, D.S. Lindsay, and C.A. Speer. 1998. "Structures of Toxoplasma Gondii Tachyzoites, Bradyzoites, and Sporozoites and Biology and Development of Tissue Cysts." *Clinical Microbiology Reviews* 11 (2): 267–99.

Dzierszinski, Florence, Marlene Mortuaire, Najoua Dendouga, Octavian Popescu, and Stanislas Tomavo. 2001. "Differential Expression of Two Plant-like Enolases with Distinct Enzymatic and Antigenic Properties during Stage Conversion of the Protozoan Parasite Toxoplasma Gondii." *Journal of Molecular Biology* 309 (5): 1017–27. doi:10.1006/jmbi.2001.4730.

Dzierszinski, Florence, Manami Nishi, Lillian Ouko, and David S Roos. 2004. "Dynamics of Toxoplasma Gondii Differentiation." *Eukaryotic Cell* 3 (4): 992–1003. doi:10.1128/EC.3.4.992.

Elliott, Ruth, Fan Li, Isabelle Dragomir, Ming Ming W Chua, Brian D Gregory, and Susan R Weiss. 2013. "Analysis of the Host Transcriptome from Demyelinating Spinal Cord of Murine Coronavirus-Infected Mice." *PLoS ONE* 8 (9): e753346. doi:10.1371/journal.pone.0075346.

Engström, Pär G, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Tyler Alioto, Jonas Behr, et al. 2013. "Systematic Evaluation of Spliced Alignment Programs for RNA-Seq Data." *Nature Methods* 10 (12): 1185–91. doi:10.1038/nmeth.2722.

Farrell, Megan, and Marc-jan Gubbels. 2014. "The Toxoplasma Gondii Kinetochore Is Required for Centrosome Association with the Centrocone (Spindle Pole)." *Cellular Microbiology* 16 (1): 78–94. doi:10.1111/cmi.12185.

Fentress, Sarah J, Michael S Behnke, Ildiko R Dunay, Mona Mashayekhi, Leah M Rommereim, Barbara A Fox, David J Bzik, et al. 2010. "Phosphorylation of Immunity-Related GTPases by a Toxoplasma Gondii-Secreted Kinase Promotes Macrophage Survival and Virulence." *Cell Host and Microbe* 8 (6): 484–95. doi:10.1016/j.chom.2010.11.005.

Fentress, Sarah J, Tobias Steinfeldt, Jonathan C Howard, and L David Sibley. 2012. "The Arginine-Rich N-Terminal Domain of ROP18 Is Necessary for Vacuole Targeting and Virulence of Toxoplasma Gondii." *Cellular Microbiology* 14 (12): 1921–33. doi:10.1111/cmi.12022.

Fiscon, Giulia, Paola Paci, and Giulio Iannello. 2015. "MONSTER v1 . 1: A Tool to Extract and

Search for RNA Non-Branching Structures." *BMC Genomics* 16: S1. doi:10.1186/1471-2164-16-S6-S1.

Fleckenstein, Martin C, Michael L Reese, Stephanie Konen-Waisman, John C. Boothroyd, Jonathan C Howard, and Tobias Steinfeldt. 2012. "A Toxoplasma Gondii Pseudokinase Inhibits Host IRG Resistance Proteins." *PLoS Biology* 10 (7): e1001358. doi:10.1371/journal.pbio.1001358.

Fox, Barbara A, Alejandra Falla, Leah M Rommereim, Tadakimi Tomita, Jason P Gigley, Corinne Mercier, Marie-France Cesbron-Delauw, Louis M Weiss, and David J Bzik. 2011. "Type II Toxoplasma Gondii KU80 Knockout Strains Enable Functional Analysis of Genes Required for Cyst Development and Latent Infection." *Eukaryotic Cell* 10 (9): 1193–1206. doi:10.1128/EC.00297-10.

Fox, Barbara A, Jessica G Ristuccia, Jason P Gigley, and David J Bzik. 2009. "Efficient Gene Replacements in Toxoplasma Gondii Strains Deficient for Nonhomologous End Joining." *Eukaryotic Cell* 8 (4): 520–29. doi:10.1128/EC.00357-08.

Fox, Barbara A, Leah M Rommereim, Rebekah B Guevara, Alejandra Falla, Myriam Andra Hortua Triana, Yanbo Sun, and David J Bzik. 2016. "The Toxoplasma Gondii Rhoptry Kinome Is Essential for Chronic Infection." *mBio* 7 (3): pii: e00193-16. doi:10.1128/mBio.00193-16.

Fritz, H, B Barr, A Packham, A Melli, and P A Conrad. 2012. "Methods to Produce and Safely Work with Large Numbers of Toxoplasma Gondii Oocysts and Bradyzoite Cysts." *Journal of Microbiological Methods* 88 (1): 47–52. doi:10.1016/j.mimet.2011.10.010.

Fritz, Heather M., Kerry R. Buchholz, Xiucui Chen, Blythe Durbin-Johnson, David M. Rocke, Patricia A. Conrad, and John C. Boothroyd. 2012. "Transcriptomic Analysis of Toxoplasma Development Reveals Many Novel Functions and Structures Specific to Sporozoites and Oocysts." *PLoS ONE* 7 (2): e29998. doi:10.1371/journal.pone.0029998.

Fritz, Heather M, Paul W Bowyer, Matthew Bogyo, Patricia A Conrad, and John C Boothroyd. 2012. "Proteomic Analysis of Fractionated Toxoplasma Oocysts Reveals Clues to Their Environmental Resistance." *PLoS ONE* 7 (1): e29955. doi:10.1371/journal.pone.0029955.

Gajria, Bindu, Amit Bahl, John Brestelli, Jennifer Dommer, Steve Fischer, Xin Gao, Mark Heiges, et al. 2008. "ToxoDB: An Integrated Toxoplasma Gondii Database Resource." *Nucleic Acids*

Research 36 (Database issue): D553-6. doi:10.1093/nar/gkm981.

Gissot, Mathieu, Krystyna A. Kelly, James W. Ajioka, John M. Greally, and Kami Kim. 2007. "Epigenomic Modifications Predict Active Promoters and Gene Structure in Toxoplasma Gondii." *PLoS Pathogens* 3 (6): e77. doi:10.1371/journal.ppat.0030077.

Graindorge, Arnault, Karine Frénal, Damien Jacot, Julien Salamun, Jean Baptiste Marq, and Dominique Soldati-Favre. 2016. "The Conoid Associated Motor MyoH Is Indispensable for Toxoplasma Gondii Entry and Exit from Host Cells." *PLoS Pathogens* 12 (1): e1005388. doi:10.1371/journal.ppat.1005388.

Grant, Gregory R, Michael H Farkas, Angel D Pizarro, Nicholas F Lahens, Jonathan Schug, Brian P Brunk, Christian J Stoeckert, John B Hogenesch, and Eric A Pierce. 2011. "Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)." *Bioinformatics* 27 (18): 2518–28. doi:10.1093/bioinformatics/btr427.

Grigg, M E, S Bonnefoy, A B Hehl, Y Suzuki, and J C Boothroyd. 2001. "Success and Virulence in Toxoplasma as the Result of Sexual Recombination Between Two Distinct Ancestries." *Science* 294 (5540): 161–65. doi:10.1126/science.1061888.

Gubbels, Marc-jan, Margaret Lehmann, Mani Muthalagi, Maria E Jerome, Carrie F Brooks, Tomasz Szatanek, Jayme Flynn, et al. 2008. "Forward Genetic Analysis of the Apicomplexan Cell Division Cycle in Toxoplasma Gondii." *PLoS Pathogens* 4 (2): e36. doi:10.1371/journal.ppat.0040036.

Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, et al. 2012. "GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74. doi:10.1101/gr.135350.111.

Hassan, Musa A, Mariane B Melo, Brian Haas, Kirk D C Jensen, and Jeroen P J Saeij. 2012. "De Novo Reconstruction of the Toxoplasma Gondii Transcriptome Improves on the Current Genome Annotation and Reveals Alternatively Spliced Transcripts and Putative Long Non-Coding RNAs." *BMC Genomics* 13: 696. doi:10.1186/1471-2164-13-696.

Hehl, Adrian B, Walter U Basso, Christoph Lippuner, Chandra Ramakrishnan, Michal Okoniewski, Robert A Walker, Michael E Grigg, Nicholas C Smith, and Peter Deplazes. 2015. "Asexual Expansion of Toxoplasma Gondii Merozoites Is Distinct from Tachyzoites and Entails

Expression of Non-Overlapping Gene Families to Attach , Invade , and Replicate within Feline Enterocytes." *BMC Genomics* 16: 66. doi:10.1186/s12864-015-1225-x.

Houseley, Jonathan, and David Tollervey. 2009. "The Many Pathways of RNA Degradation." *Cell* 136 (4): 763–76. doi:10.1016/j.cell.2009.01.019.

Hout, Michael C, Megan H Papesh, and Stephen D Goldinger. 2013. "Multidimensional Scaling." *Wiley Interdisciplinary Reviews: Cognitive Science* 4 (1): 93–103. doi:10.1002/wcs.1203.

Howe, Daniel K, and L David Sibley. 1995. "Toxoplasma Gondii Comprises Three Clonal Lineages: Correlation of Parasite Genotype with Human Disease." *Journal of Infectious Diseases* 172 (6): 1561–66.

Huynh, My-Hang, Martin J Boulanger, and Vern B Carruthers. 2014. "A Conserved Apicomplexan Microneme Protein Contributes to Toxoplasma Gondii Invasion and Virulence." *Infection and Immunity* 82 (10): 4358–68. doi:10.1128/IAI.01877-14.

Huynh, My-Hang, and Vern B Carruthers. 2009. "Tagging of Endogenous Genes in a Toxoplasma Gondii Strain Lacking Ku80." *Eukaryotic Cell* 8 (4): 530–39. doi:10.1128/EC.00358-08.

Huynh, My-Hang, and Vern B Carruthers. 2016. "A Toxoplasma Gondii Ortholog of Plasmodium GAMA Contributes to Parasite Attachment and Cell Invasion." *mSphere* 1 (1): pii: e00012-16. doi:10.1128/mSphere.00012-16.

Incarnato, Danny, and Salvatore Oliviero. 2016. "The RNA Epistructurome: Uncovering RNA Function by Studying Structure and Modi Fi Cations." *Trends in Biotechnology*, pii: S0167-7799(16)30203-7. doi:10.1016/j.tibtech.2016.11.002.

Jain, Miten, Hugh E Olsen, Benedict Paten, and Mark Akeson. 2016. "The Oxford Nanopore MinION: Delivery of Nanopore Sequencing to the Genomics Community." *Genome Biology* 17 (1): 239. doi:10.1186/s13059-016-1103-0.

Juránková, Jana, Marieke Opsteegh, Helena Neumayerová, Kamil Kovarcík, Anita Frencová, Vojtech Baláz, Jirí Volf, and Bretislav Koudela. 2013. "Quantification of Toxoplasma Gondii in Tissue Samples of Experimentally Infected Goats by Magnetic Capture and Real-Time PCR." *Veterinary Parasitology* 193 (1–3): 95–99. doi:10.1016/j.vetpar.2012.11.016.

Katz, Yarden, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. 2010. "Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation." *Nature*

Methods 7 (12): 1009–15. doi:10.1038/nmeth.1528.

Kehr, Sebastian, Nicole Sturm, Stefan Rahlfs, Jude M Przyborski, and Katja Becker. 2010. "Compartmentation of Redox Metabolism in Malaria Parasites." *PLoS Pathogens* 6 (12): e1001242. doi:10.1371/journal.ppat.1001242.

Kent, W James. 2002. "BLAT — The BLAST -Like Alignment Tool." *Genome Research* 12 (4): 656–64. doi:10.1101/gr.229202.

Khan, Asis, Natalie Miller, David S. Roos, J.P. Dubey, Daniel Ajzenberg, Marie Laure Darde, James W. Ajioka, Benjamin Rosenthal, and L. David Sibley. 2011. "A Monomorphic Haplotype of Chromosome Ia Is Associated with Widespread Success in Clonal and Nonclonal Populations of Toxoplasma Gondii." *mBio* 2 (6): e00228-11. doi:10.1128/mBio.00228-11.

Kim, Eun-Deok, and Sibum Sung. 2012. "Long Noncoding RNA: Unveiling Hidden Layer of Gene Regulatory Networks." *Trends in Plant Science* 17 (1): 16–21. doi:10.1016/j.tplants.2011.10.008.

Kim, Kami, and Louis M Weiss. 2004. "Toxoplasma Gondii: The Model Apicomplexan." *International Journal for Parasitology* 34 (3): 423–32. doi:10.1016/j.ijpara.2003.12.009.

Kissinger, Jessica C, Bindu Gajria, Li Li, Ian T Paulsen, and David S Roos. 2003. "ToxoDB: Accessing the Toxoplasma Gondii Genome." *Nucleic Acids Research* 31 (1): 234–36. doi:10.1093/nar/gkg072.

Köhler, Sabine, Charles F Delwiche, Paul W Denny, Lewis G Tilney, Paul Webster, R J M Wilson, Jeffrey D Palmer, and David S Roos. 1997. "A Plastid of Probable Green Algal Origin in Apicomplexan Parasites." *Science* 275 (5305): 1485–90.

Kornblihtt, Alberto R, Ignacio E Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J Muñoz. 2013. "Alternative Splicing: A Pivotal Step between Eukaryotic Transcription and Translation." *Nature Reviews. Molecular Cell Biology* 14 (3): 153–66. doi:10.1038/nrm3525.

Lamesch, Philippe, Tanya Z Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, et al. 2012. "The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools." *Nucleic Acids Research* 40 (Database Issue): D1202-10. doi:10.1093/nar/gkr1090.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *BMC Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Leroux, Louis-philippe, Dayal Dasanayake, Leah M Rommereim, Barbara A Fox, David J Bzik, Armando Jardim, and Florence S Dzierszinski. 2015. “Secreted Toxoplasma Gondii Molecules Interfere with Expression of MHC-II in Interferon Gamma-Activated Macrophages.” *International Journal for Parasitology* 45 (5): 319–32. doi:10.1016/j.ijpara.2015.01.003.

Li, Fan, Qi Zheng, Paul Ryvkin, Isabelle Dragomir, Yaanik Desai, Subhadra Aiyer, Otto Valladares, et al. 2012. “Global Analysis of RNA Secondary Structure in Two Metazoans.” *CellReports* 1 (1): 69–82. doi:10.1016/j.celrep.2011.10.002.

Li, Fan, Qi Zheng, Lee E Vandivier, Matthew R Willmann, Ying Chen, and Brian D Gregory. 2012. “Regulatory Impact of RNA Secondary Structure across the Arabidopsis Transcriptome.” *The Plant Cell* 24: 4346–59. doi:10.1105/tpc.112.104232.

Liao, Qi, Changning Liu, Xiongying Yuan, Shuli Kang, Ruoyu Miao, Hui Xiao, Guoguang Zhao, et al. 2011. “Large-Scale Prediction of Long Non-Coding RNA Functions in a Coding – Non-Coding Gene Co-Expression Network.” *Nucleic Acids Research* 39 (9): 3864–78. doi:10.1093/nar/gkq1348.

Lorenzi, Hernan, Asis Khan, Michael S. Behnke, Sivaranjani Namasivayam, Lakshmipuram S. Swapna, Michalis Hadjithomas, Svetlana Karamycheva, et al. 2016. “Local Admixture of Amplified and Diversified Secreted Pathogenesis Determinants Shapes Mosaic Toxoplasma Gondii Genomes.” *Nature Communications* 7: 10147. doi:10.1038/ncomms10147.

Martin, Jeffrey a., and Zhong Wang. 2011. “Next-Generation Transcriptome Assembly.” *Nature Reviews Genetics* 12 (10): 671–82. doi:10.1038/nrg3068.

Matrajt, Mariana, Craig D Platt, Anurag D Sagar, a Lindsay, C Moulton, and David S Roos. 2004. “Transcript Initiation, Polyadenylation, and Functional Promoter Mapping for the Dihydrofolate Reductase-Thymidylate Synthase Gene of Toxoplasma Gondii.” *Molecular and Biochemical Parasitology* 137 (2): 229–38. doi:10.1016/j.molbiopara.2003.12.015.

McFadden, Geoffrey Ian, and Ellen Yeh. 2017. “The Apicoplast: Now You See It, Now You Don’t.” *International Journal for Parasitology* 47 (2–3): 137–44. doi:10.1016/j.ijpara.2016.08.005.

McManus, Michael T, and Phillip A Sharp. 2002. “Gene Silencing in Mammals by Small Interfering RNAs.” *Nature Reviews Genetics* 3 (10): 737–47. doi:10.1038/nrg908.

Meissner, Markus, Manuela S Breinich, Paul R Gilson, and Brendan S Crabb. 2007. “Molecular Genetic Tools in Toxoplasma and Plasmodium: Achievements and Future Needs.” *Current Opinion in Microbiology* 10 (4): 349–56. doi:10.1016/j.mib.2007.07.006.

Meissner, Markus, and Dominique Soldati. 2005. “The Transcription Machinery and the Molecular Toolbox to Control Gene Expression in Toxoplasma Gondii and Other Protozoan Parasites.” *Microbes and Infection* 7 (13): 1376–84. doi:10.1016/j.micinf.2005.04.019.

Minot, S., M. B. Melo, F. Li, D. Lu, W. Niedelman, S. S. Levine, and J. P. J. Saeij. 2012. “Admixture and Recombination among Toxoplasma Gondii Lineages Explain Global Genome Diversity.” *Proceedings of the National Academy of Sciences* 109 (33): 13458–63. doi:10.1073/pnas.1117047109.

Mu, Xinmeng Jasmine, Zhi John Lu, Yong Kong, Hugo Y K Lam, and Mark B Gerstein. 2011. “Analysis of Genomic Variation in Non-Coding Elements Using Population-Scale Sequencing Data from the 1000 Genomes Project.” *Nucleic Acids Research* 39 (16): 7058–76. doi:10.1093/nar/gkr342.

Mudge, Jonathan M, and Jennifer Harrow. 2016. “The State of Play in Higher Eukaryote Gene Annotation.” *Nature Reviews Genetics* 17 (12): 758–72. doi:10.1038/nrg.2016.119.

Nishikura, Kazuko. 2001. “A Short Primer on RNAi: RNA-Directed RNA Polymerase Acts as a Key Catalyst.” *Cell* 107 (4): 415–18.

Niu, Meng, Ehsan S Tabari, and Zhengchang Su. 2014. “De Novo Prediction of Cis-Regulatory Elements and Modules through Integrative Analysis of a Large Number of ChIP Datasets.” *BMC Genomics* 15: 1047. doi:10.1186/1471-2164-15-1047.

O’Connell, Ryan M, Dinesh S Rao, and David Baltimore. 2012. “microRNA Regulation of Inflammatory Responses.” *Annual Review of Immunology* 30: 295–312. doi:10.1146/annurev-immunol-020711-075013.

Oliveira Dal’Molin, Cristiana G de, Camila Orellana, Leigh Gebbie, Jennifer Steen, Mark P Hodson, Panagiotis Chrysanthopoulos, Manuel R Plan, Richard McQualter, Robin W. Palfreyman, and Lars K Nielsen. 2016. “Metabolic Reconstruction of Setaria Italica: A Systems Biology Approach for Integrating Tissue-Specific Omics and Pathway Analysis of

Bioenergy Grasses." *Frontiers in Plant Science* 7: 1138. doi:10.3389/fpls.2016.01138.

Patil, Veerupaxagouda, Pamela J. Lescault, Dario Lirussi, Ann B. Thompson, and Mariana Matrajt. 2013. "Disruption of the Expression of a Non-Coding RNA Significantly Impairs Cellular Differentiation in Toxoplasma Gondii." *International Journal of Molecular Sciences* 14 (1): 611–24. doi:10.3390/ijms14010611.

Pfefferkorn, E.R., Lorraine C. Pferfferkorn, and Emerson D. Colby. 1977. "Development of Gametes and Oocysts in Cats Fed Cysts Derived from Cloned Trophozoites of Toxoplasma Gondi." *The Journal of Parasitology* 63 (1): 158–59.

Pittman, Kelly J, Matthew T Aliota, and Laura J Knoll. 2014. "Dual Transcriptional Profiling of Mice and Toxoplasma Gondii during Acute and Chronic Infection." *BMC Genomics* 15: 806. doi:10.1186/1471-2164-15-806.

Plaschka, C, L Lariviere, L Wenzeck, M Seizl, M Hemann, D Tegunov, E V Petrotchenko, et al. 2015. "Architecture of the RNA Polymerase II – Mediator Core Initiation Complex." *Nature* 518 (7539): 376–80. doi:10.1038/nature14229.

Radhakrishnan, Aditya, and Rachel Green. 2016. "Connections Underlying Translation and mRNA Stability." *Journal of Molecular Biology* 428 (18): 3558–64. doi:10.1016/j.jmb.2016.05.025.

Radke, Jay R, Michael S Behnke, Aaron J Mackey, Josh B Radke, David S Roos, and Michael W White. 2005. "The Transcriptome of Toxoplasma Gondii." *BMC Biology* 3: 26. doi:10.1186/1741-7007-3-26.

Raina, Medha, and Michael Ibba. 2014. "tRNAs as Regulators of Biological Processes." *Frontiers in Genetics* 5: 171. doi:10.3389/fgene.2014.00171.

Ralph, Stuart A, Giel G Van Dooren, Ross F Waller, Michael J Crawford, Martin J Fraunholz, Bernardo J Foth, Christopher J Tonkin, David S Roos, and Geoffrey I McFadden. 2004. "Tropical Infectious Diseases: Metabolic Maps and Functions of the Plasmodium Falciparum Apicoplast." *Nature Reviews Microbiology* 2 (3): 203–16. doi:10.1038/nrmicro843.

Reese, Michael L, Gusti M Zeiner, Jeroen P J Saeij, John C Boothroyd, and Jon P Boyle. 2011. "Polymorphic Family of Injected Pseudokinases Is Paramount in Toxoplasma Virulence." *Proceedings of the National Academy of Sciences* 108 (23): 9625–30. doi:10.1073/pnas.1015980108.

Reid, Adam James, Sarah J Vermont, James A Cotton, David Harris, Grant A Hill-Cawthorne, Stephanie Konen-Waisman, Sophia M Latham, et al. 2012. “Comparative Genomics of the Apicomplexan Parasites Toxoplasma Gondii and Neospora Caninum: Coccidia Differing in Host Range and Transmission Strategy.” *PLoS Pathogens* 8 (3): e1002567. doi:10.1371/journal.ppat.1002567.

Rinn, John L., Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha A. Brugmann, L. Henry Goodnough, et al. 2007. “Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs.” *Cell* 129 (7): 1311–23. doi:10.1016/j.cell.2007.05.022.

Roiko, Marijo S, Nadezhda Svezhova, and Vern B Carruthers. 2014. “Acidification Activates Toxoplasma Gondii Motility and Egress by Enhancing Protein Secretion and Cytolytic Activity.” *PLoS Pathogens* 10 (11): e10004488. doi:10.1371/journal.ppat.1004488.

Rommereim, Leah M, Miryam A Hortua Triana, Alejandra Falla, Kiah L Sanders, Rebekah B Guevara, David J Bzik, and Barbara A Fox. 2013. “Genetic Manipulation in Δku80 Strains for Functional Genomic Analysis of Toxoplasma Gondii.” *Journal of Visualized Experiments* (77): e50598. doi:10.3791/50598.

Roos, David S, Robert G K Donald, Naomi S Morrissette, and A Lindsay C Moulton. 1995. “Molecular Tools for Genetic Dissection of the Protozoan Parasite Toxoplasma Gondii.” In *Methods in Cell Biology*, 45:27–63. Elsevier.

Rosowski, Emily E, Diana Lu, Lindsay Julien, Lauren Rodda, Rogier A Gaiser, Kirk D C Jensen, and Jeroen P J Saeij. 2011. “Strain-Specific Activation of the NF-kappaB Pathway by GRA15, a Novel Toxoplasma Gondii Dense Granule Protein.” *The Journal of Experimental Medicine* 208 (1): 195–212. doi:10.1084/jem.20100717.

Saeij, J P J, S Coller, J P Boyle, M E Jerome, M W White, and J C Boothroyd. 2007. “Toxoplasma Co-Opts Host Gene Expression by Injection of a Polymorphic Kinase Homologue.” *Nature* 445 (7125): 324–27. doi:10.1038/nature05395.

Salamov, Asaf A, and Victor V Solovyev. 2000. “Ab Initio Gene Finding in Drosophila Genomic DNA.” *Genome Research* 10 (4): 516–22.

Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. “A Single-Molecule Long-Read Survey of the Human Transcriptome.” *Nature Biotechnology* 31 (11): 1009–14.

doi:10.1038/nbt.2705.

Shen, Shihao, Juw Won Park, Zhi-xiang Lu, Lan Lin, Michael D Henry, Ying Nian Wu, and Qing Zhou. 2014. “rMATS : Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data.” *Proceedings of the National Academy of Sciences* 111 (51): E5593-5601. doi:10.1073/pnas.1419161111.

Sibley, L David, Marinella Messina, and Ingrid R Niesman. 1994. “Stable DNA Transformation in the Obligate Intracellular Parasite Toxoplasma Gondii by Complementation of Tryptophan Auxotrophy.” *Proceedings of the National Academy of Sciences* 91 (12): 5508–12.

Sibley, L David, Dana G Mordue, Chunlei Su, Paul M Robben, and Dan K Howe. 2002. “Genetic Approaches to Studying Virulence and Pathogenesis in Toxoplasma Gondii.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 357 (1417): 81–88. doi:10.1098/rstb.2001.1017.

Sidik, Saima M., Diego Huet, Suresh M. Ganesan, My-Hang Huynh, Tim Wang, Armiyaw S. Nasamu, Prathapan Thiru, et al. 2016. “A Genome-Wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes.” *Cell* 166 (6): 1423–1435.e12. doi:10.1016/j.cell.2016.08.019.

Singh, Upinder, Jeremy L Brewer, and John C Boothroyd. 2002. “Genetic Analysis of Tachyzoite to Bradyzoite Differentiation Mutants in Toxoplasma Gondii Reveals a Hierarchy of Gene Induction.” *Molecular Microbiology* 44 (3): 721–33.

Soete, M, Fortier D. Camus, and J. F. Dubremetz. 1993. “Toxoplasma Gondii: Kinetics of Bradyzoite-Tachyzoite Interconversion in Vitro.” *Experimental Parasitology* 76 (3): 259–64.

Soldati, Dominique. 1996. “Molecular Genetic Strategies in Toxoplasma Gondii: Close in on a Successful Invader.” *FEBS Letters* 389 (1): 80–83.

Sullivan, William J, Joshua B Radke, Kami Kim, and Michael W White. 2013. *Epigenetic and Genetic Factors That Regulate Gene Expression in Toxoplasma Gondii*. *Toxoplasma Gondii: The Model Apicomplexan - Perspectives and Methods*. Second Edi. Elsevier. doi:10.1016/B978-0-12-396481-6.00018-0.

Taganov, Konstantin D, Mark P Boldin, Kuang-jung Chang, and David Baltimore. 2006. “NF-kappaB-Dependent Induction of microRNA miR-146, an Inhibitor Targeted to Signaling Proteins of Innate Immune Responses.” *Proceedings of the National Academy of Sciences*

103 (33): 12481–86. doi:10.1073/pnas.0605298103.

Tenter, Astrid M, Anja R Heckeroth, and Louis M Weiss. 2000. “Toxoplasma Gondii : From Animals to Humans.” *International Journal for Parasitology* 30 (12–13): 1217–58.

Todeschini, Anne Laure, Adrien Georges, and Reiner A. Veitia. 2014. “Transcription Factors: Specific DNA Binding and Specific Gene Regulation.” *Trends in Genetics* 30 (6): 211–19. doi:10.1016/j.tig.2014.04.002.

Trapnell, Cole, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. 2013. “Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq.” *Nature Biotechnology* 31 (1): 46–53. doi:10.1038/nbt.2450.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15. doi:10.1038/nbt.1621.

Turner, Anne-marie W, and Kevin V Morris. 2010. “Controlling Transcription with Noncoding RNAs in Mammalian Cells.” *Biotechniques* 48 (6): ix–xvi. doi:10.2144/000113442.

Uhlén, Mathias, Linn Fagerberg, Björn M Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. “Proteomics. Tissue-Based Map of the Human Proteome.” *Science* 347 (6220): 1260419. doi:10.1126/science.1260419.

Vaquero-Garcia, Jorge, Alejandro Barrera, Matthew R Gazzara, Juan Gonzales-Vallinas, Nicholas F Lahens, John B Hogenesch, Kristen W Lynch, and Yoseph Barash. 2016. “A New View of Transcriptome Complexity and Regulation through the Lens of Local Splicing Variations.” *eLIFE* 5: e11752. doi:10.7554/eLife.11752.

Venables, Julian P, Roscoe Klinck, Anne Bramard, Lyna Inkel, Genevieve Dufresne-Martin, Chushin Koh, Julien Gervais-bird, et al. 2008. “Identification of Alternative Splicing Markers for Breast Cancer.” *Cancer Research* 68 (22): 9525–31. doi:10.1158/0008-5472.CAN-08-1769.

Walker, Robert, Mathieu Gissot, Matthew M Croken, Ludovic Huot, David Hot, Kami Kim, and Stanislas Tomavo. 2013. “The Toxoplasma Nuclear Factor TgAP2XI-4 Controls Bradyzoite Gene Expression and Cyst Formation.” *Molecular Microbiology* 87 (3): 641–55. doi:10.1111/mmi.12121.

Walker, Robert, Mathieu Gissot, Ludovic Huot, Tchilabalo Dilezitoko Alayi, David Hot, Guillemette Marot, Christine Schaeffer-Reiss, Alain Van Dorsselaer, Kami Kim, and Stanislas Tomavo. 2013. “Toxoplasma Transcription Factor TgAP2XI-5 Regulates the Expression of Genes Involved in Parasite Virulence and Host Invasion.” *Journal of Biological Chemistry* 288 (43): 31127–38. doi:10.1074/jbc.M113.486589.

Wang, Eric T, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, and Lu Zhang. 2008. “Alternative Isoform Regulation in Human Tissue Transcriptomes.” *Nature* 456 (7221): 470–76. doi:10.1038/nature07509.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics* 10 (1): 57–63. doi:10.1038/nrg2484.

Wasmuth, James D., Viviana Pszenny, Simon Haile, Emily M. Jansen, Alexandra T. Gast, Alan Sher, Jon P. Boyle, Martin J. Boulanger, John Parkinson, and Michael E. Grigg. 2012. “Integrated Bioinformatic and Targeted Deletion Analyses of the SRS Gene Superfamily Identify SRS29C as a Negative Regulator of Toxoplasma Virulence.” *mBio* 3 (6): pii: e00321-12. doi:10.1128/mBio.00321-12.

Wastling, J M, D Xia, a Sohal, M Chaussepied, a Pain, and G Langsley. 2009. “Proteomes and Transcriptomes of the Apicomplexa--Where’s the Message?” *International Journal for Parasitology* 39 (2): 135–43. doi:10.1016/j.ijpara.2008.10.003.

Wei, Chaochun, and Michael R Brent. 2006. “Using ESTs to Improve the Accuracy of de Novo Gene Prediction.” *BMC Bioinformatics* 7: 327. doi:10.1186/1471-2105-7-327.

Westermann, Alexander J, Stanislaw A Gorski, and Jörg Vogel. 2012. “Dual RNA-Seq of Pathogen and Host.” *Nature Reviews Microbiology* 10 (9): 618–30. doi:10.1038/nrmicro2852.

Wong, Koon Ho, Yi Jin, and Kevin Struhl. 2014. “TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape.” *Molecular Cell* 54 (4): 601–12. doi:10.1016/j.molcel.2014.03.024.

Wu, Thomas D, and Serban Nacu. 2010. “Fast and SNP-Tolerant Detection of Complex Variants and Splicing in Short Reads.” *Bioinformatics* 26 (7): 873–81. doi:10.1093/bioinformatics/btq057.

Xia, Dong, Sanya J Sanderson, Andrew R Jones, Judith H Prieto, John R Yates, Elizabeth

Bromley, Fiona M Tomley, et al. 2008. “The Proteome of Toxoplasma Gondii: Integration with the Genome Provides Novel Insights into Gene Expression and Annotation.” *Genome Biology* 9 (7): R116. doi:10.1186/gb-2008-9-7-r116.

Zeiner, Gusti M, Kara L Norman, J Michael Thomson, Scott M Hammond, and John C Boothroyd. 2010. “Toxoplasma Gondii Infection Specifically Increases the Levels of Key Host microRNAs.” *PLoS ONE* 5 (1): e8742.